Prototype Learning for Out-of-Distribution Polyp Segmentation

Nikhil Kumar Tomar[1], **Debesh Jha**[2], **Ulas Bagci**[1]

[1]Northwestern University
[2]University of South Dakota
nikhilroxtomar@gmail.com, debesh.jha@usd.edu, ulas.bagci@northwestern.edu

Abstract

Existing polyp segmentation models from colonoscopy images often fail to provide reliable segmentation results on datasets from different centers, limiting their applicability. Our objective in this study is to create a robust and well-generalized segmentation model named *PrototypeLab* that can assist in polyp segmentation. To achieve this, we incorporate various lighting modes such as White light imaging (WLI), Blue light imaging (BLI), Linked color imaging (LCI), and Flexible spectral imaging color enhancement (FICE) into our new segmentation model, which learns to create prototypes for each class of object present in the images. These prototypes represent the characteristic features of the objects, such as their shape, texture, and color. Our model is designed to perform effectively on out-of-distribution (OOD) datasets from multiple centers. We first generate a coarse mask that is used to learn prototypes for the main object class, which are then employed to generate the final segmentation mask. By using prototypes to represent the main class, our approach handles the variability present in the medical images and generalizes well to new data since prototypes capture the underlying distribution of the data. PrototypeLab offers a promising solution with a dice coefficient of $\geq$ 90% and mIoU $\geq$ 85% with a near real-time processing speed for polyp segmentation. It achieved superior performance on OOD datasets compared to 16 state-of-the-art image segmentation architectures, potentially improving clinical outcomes. Codes will be made available at https://github.com/nikhilroxtomar/PrototypeLab.

Introduction

The Cancer statistic 2023 (Siegel et al. 2023) estimates that colorectal cancer (CRC)will be the third leading cause of cancer-related incidence and death in the United States. The United States Preventive Services Task Force (USPSTF) recommends CRC screening at 45 years of age (Ng, P. May, and Schrag 2021). Thus, regular screening is essential as CRC does not show symptoms (bleeding in the stool, constipation or diarrhoea) at an early stage. Studies have shown the lesion miss rate to be 26% (Corley et al. 2014). This emphasizes the need for an accurate and reliable screening CADx method for reducing the polyp miss-rate and contributing to the reduction of CRC related death.

Encoder-decoder based networks are widely used for automatic polyp segmentation (Dong et al. 2021; Fan et al. 2020; Jha et al. 2020a; Rahman and Marculescu 2023; Sanderson and Matuszewski 2022; Tang et al. 2022; Tomar et al. 2022; Zhang, Liu, and Hu 2021; Zhao, Zhang, and Lu 2021). Dong et al. (Dong et al. 2021) proposed PolypPVT that is based on pyramid vision transformer for automatic polyp segmentation. Wang et al. (Wang et al. 2022) proposed SSFormer, which uses a pyramid Transformer encoder and progressive locality decoder to improve the performance of polyp segmentation. Encoder-decoder based methods achieved improved accuracy on regular or large-sized polyps. However, most methods are developed on white light imaging modality and are tested on an in-distribution dataset (tested on the same center). They readily fail on OOD datasets such as small, diminutive, flat, sessile, or partially visible polyps and in the presence of camouflage and noisy images. Moreover, the complex morphological structures, indistinct boundaries between polyps and mucosa, and varying sizes make polyp segmentation more challenging. Therefore, there is an urgent need to develop a more generalizable and robust polyp segmentation method.

We hypothesize that a prototype based segmentation can be a strong approach for OOD generalization because it relies on the creation of prototypes that capture the essential features of each class of objects in the images. These prototypes can be used to identify and segment objects even in images that differ significantly from those used during training. An algorithm performing well on OOD dataset could prevent models from making inaccurate diagnoses or treatment planning. Failure to do so can result in false positives or false negatives leading to misdiagnosis, which might have adverse consequences for the patient. Therefore, the model must perform well on OOD datasets to be useful in clinical settings. In this work, we evaluate the proposed model on three datasets collected in different countries captured under different conditions and image enhancement techniques, including OOD datasets to study the generalization ability of the model, which is critical for the development of CADx system.

Summary of our contributions: (1) We propose to develop a *prototype learning* algorithm for medical image segmentation. To generate multiple prototypes, we design a Prototype Generation Module (PGM) by capturing the under-

lying data distribution. These prototypes handle variability and exhibit strong generalization capabilities when applied to novel data, **(2)** We present a *Coarse Mask Generation Module (CMGM)* to improve the accuracy, efficiency, and generalization capabilities of the polyp segmentation network, **(3)** We encourage our network to operate on multi-scale fashion through an *Encoder Feature Fusion Module (EFFM)*, **(4)** We devise a *Prototype Mask Generation Module (PMGM)* to generate a final prototype mask using the prototypes generated by the PGM and the output feature map of the EFFM, and **(5)** we conduct a thorough analysis and evaluation on multi-center datasets and perform an OOD generalization test on different datasets. Our experiments and results show that PrototypeLab outperforms 16 SOTA medical image segmentation methods on three publicly available polyp datasets in terms of accuracy and efficiency.

Method

The proposed PrototypeLab is a new image segmentation architecture with integrated *prototype learning*, generating high-quality segmentation masks. PrototypeLab consists of five key components: Pyramid Vision Transformer (PVT) encoder, Coarse Mask Generation Module (CMGM), Prototype Generation Module (PGM), Decoder, and Prototype Mask Generation Module (PMGM) (Figure 1). The CMGM, PGM, and PMGM are integral components that effectively contribute to the prototype learning process and collectively enhance the overall performance of the proposed architecture.

Pyramid Vision Transformer (PVT) Encoder

In PrototypeLab, the pyramid vision transformer (PVT) is used as a pre-trained encoder, extracting multi-scale features from an input image. An input image $I \in \mathbb{R}^{H\times W\times 3}$ is fed to the PVT-encoder to obtain four distinct feature maps $X_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where $i \in \{1,2,3,4\}$ and $C_i \in \{64,128,320,512\}$. These feature maps are then used in the other modules of PrototypeLab.

Coarse Mask Generation Module (CMGM)

The CMGM module generates a loose mask using features learned by the PVT encoder. It begins with an upsampling of X_4 to double its spatial dimensions, followed by concatenation with X_3 and passing through 3×3 Conv2D-BN-ReLU layers. The module incorporates a novel *Large Kernel Dilated Convolution (LKDC)* block (Figure 2) consisting of two parallel convolution layers with large kernel size of 7×7 and 13×13 factorized into $\{7 \times 1\}\{1 \times 7\}$ and $\{13 \times 1\}\{1 \times 13\}$ respectively. The outputs of these layers are concatenated and fed into parallel dilated convolution layers with dilation rates of $r = 1, 2, 4$. The utilization of large kernel size along dilated convolutions enhance the receptive field for capturing contextual information and handling objects of varying scales. The output is processed through a 1×1 Conv2D-BN-ReLU layer and upsampled by a factor of two. A final 1×1 convolution layer with sigmoid activation generates the coarse mask. The LKDC block enhances accuracy, efficiency, and generalization capabilities of the segmentation network.

Prototype Generation Module (PGM)

The PGM module utilizes PVT encoder features and a coarse mask to generate multiple prototypes. It applies 3×3 Conv2D-BN-ReLU layers to the PVT-encoder feature X_i. The feature maps are then element-wise multiplied with the coarse mask, followed by mask average pooling. This process yields four prototypes $P = \{p_1, p_2, p_3, p_4\}$ from $X_i = \{X_1, X_2, X_3, X_4\}$. Learning multiple prototypes allows for a richer representation of diverse visual patterns. By utilizing different features from the PVT encoder, we can capture a wider range of information and enhance the model's ability to recognize various objects and scenarios. This approach improves the robustness of the system by increasing its adaptability to different input conditions, resulting in more accurate and reliable predictions. The number of prototypes can be adjusted based on the problem complexity.

For multi-scale feature fusion of PVT encoder features, we introduce the Encoder Feature Fusion Module (EFFM) (Figure 2). It involves upsampling X_4 and concatenating it with X_3, followed by 3×3 Conv2D-BN-ReLU layers. This process is repeated for upsampling and concatenation with X_2, and again with X_1. Four parallel convolution layers are used, including a 1×1 convolution layer and three factorized convolution layers with dilated convolutions. The outputs of these layers are concatenated and passed through a final 1×1 convolution layer to generate the final prototype p_5.

Decoder

In the decoder, the upsampled feature map from the CMGM is first concatenated with X_2 and passed through a residual block. Then, the feature map is upsampled by a factor of two and concatenated with X_1, followed by another residual block. Subsequently, we upsample the feature map to increase the spatial dimensions by a factor of four, and then concatenate it with the original input image. This is followed by another residual block. Finally, the output of this block is used to generate prototype masks.

Prototype Mask Generation Module and Final Mask

The PMGM utilizes cosine similarity to generate five sets of masks by comparing the decoder's output feature map with multiple prototypes from the PGM. To create the final segmentation mask, we concatenate the decoder's output feature map with the prototype masks. This concatenated feature map undergoes a residual block, followed by a 1×1 convolution and a sigmoid activation function. The proposed framework effectively incorporates the CMGM, PGM, decoder, and PMGM to generate accurate and efficient polyp segmentation masks.

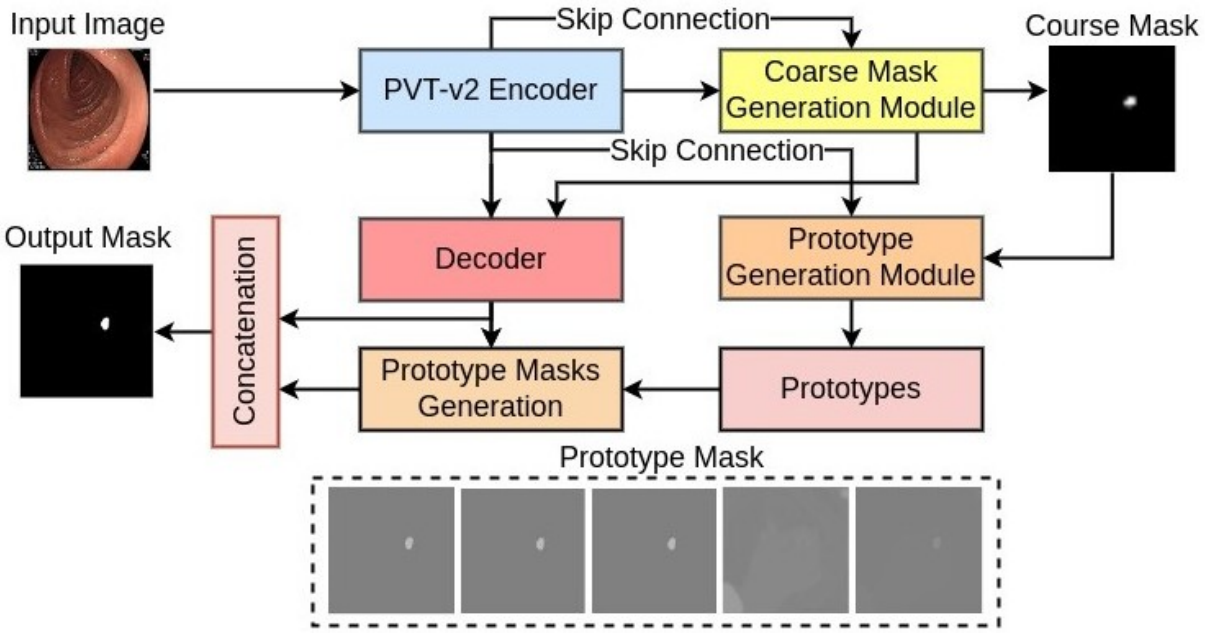

Figure 1: The overall architecture of the proposed PrototypeLab. The input image is fed to the PVT-encoder to generate a coarse mask, which is combined with encoder features in the prototype generation module to generate various prototypes. Subsequently, the decoder is employed, which produces a feature map that is used to create the prototype masks. Finally, the 'output mask' is generated.

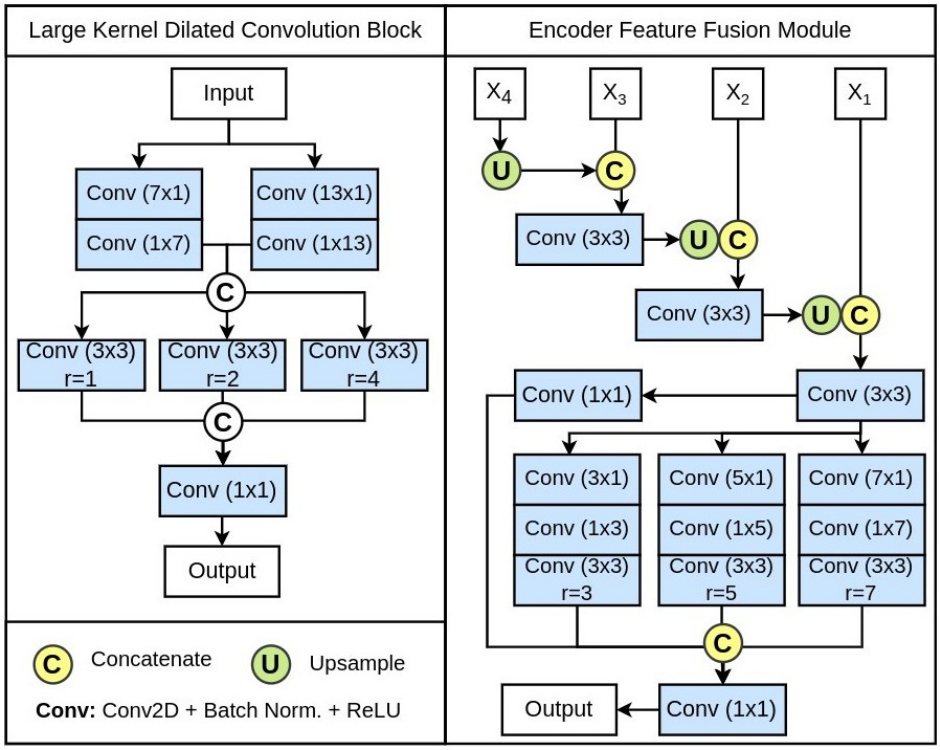

Figure 2: The diagram of the Large Kernel Dilated Convolution block and Encoder Feature Fusion Module.

Experiments and Results

Datasets: We use BKAI-IGH (Lan et al. 2021), Kvasir-SEG (Jha et al. 2020b), and PolypGen (Ali et al. 2023) dataset for experimentation. The BKAI-IGH consists of 1000 images with the corresponding ground truth. The

Table 1: Result of models trained and tested on BKAI-IGH (Lan et al. 2021). 'Red', 'Green' and 'Blue' colors represent the highest, second highest and third highest scores.

Method	mDSC	mIoU	Recall	Precision	F2	HD
U-Net (Ronneberger, Fischer, and Brox 2015)	0.8286	0.7599	0.8295	0.8999	0.8264	3.17
DeepLabV3+ (Chen et al. 2018)	0.8938	0.8314	0.8870	0.9333	0.8882	2.90
PraNet (Fan et al. 2020)	0.8904	0.8264	0.8901	0.9247	0.8885	2.94
MSNet (Zhao, Zhang, and Lu 2021)	0.9013	0.8402	0.8868	0.9423	0.8913	2.85
TransFuse-S (Zhang, Liu, and Hu 2021)	0.8599	0.7819	0.8531	0.9075	0.8530	3.04
TransFuse-L (Zhang, Liu, and Hu 2021)	0.8747	0.8105	0.8736	0.9235	0.8723	2.96
Polyp-PVT (Dong et al. 2021)	0.8995	0.8379	0.9016	0.9238	0.8986	2.88
UACANet (Kim, Lee, and Kim 2021)	0.8945	0.8275	0.8870	0.9297	0.8882	2.86
DuAT (Tang et al. 2022)	0.9140	0.8563	0.9038	0.9437	0.9066	2.77
CaraNet (Lou et al. 2022)	0.8962	0.8329	0.8939	0.9273	0.8937	2.91
SSFormer-S (Wang et al. 2022)	0.9111	0.8527	0.9043	0.9391	0.9060	2.81
SSFormer-L (Wang et al. 2022)2022	0.9124	0.8508	0.9005	0.9400	0.9041	2.74
UNeXt (Valanarasu and Patel 2022)	0.4758	0.3797	0.5814	0.5820	0.5132	4.49
LDNet (Zhang et al. 2022)	0.8927	0.8254	0.8867	0.9153	0.8874	2.94
TGANet (Tomar et al. 2022)	0.9023	0.8409	0.9025	0.9208	0.9002	2.84
PVT-Cascade (Rahman and Marculescu 2023)	0.9123	0.8534	**0.9223**	0.9212	0.9167	2.81
PrototypeLab	**0.9243**	**0.8744**	0.9194	**0.9494**	**0.9202**	**2.70**

Table 2: Quantitative results of the model trained and tested on Kvasir-SEG (Jha et al. 2020b).

Method	mDSC	mIoU	Recall	Precision	F2	HD
U-Net (Ronneberger, Fischer, and Brox 2015)	0.8264	0.7472	0.8503	0.8703	0.8352	4.57
DeepLabV3+ (Chen et al. 2018)	0.8837	0.8172	0.9014	0.9028	0.8900	4.10
PraNet (Fan et al. 2020)	0.8943	0.8296	0.9060	0.9126	0.8976	4.00
MSNet (Zhao, Zhang, and Lu 2021)	0.8859	0.8217	0.9006	0.9110	0.8901	4.01
TransFuse-S (Zhang, Liu, and Hu 2021)	0.8780	0.8079	0.8898	0.9090	0.8813	4.09
TransFuse-L (Zhang, Liu, and Hu 2021)	0.8768	0.8115	0.8842	0.9198	0.8771	4.05
Polyp-PVT (Dong et al. 2021)	0.8960	0.8328	**0.9440**	0.8811	0.9164	3.91
UACANet (Kim, Lee, and Kim 2021)	0.8835	0.8133	0.9085	0.8947	0.8937	4.19
DuAT (Tang et al. 2022)	0.8903	0.8294	0.9186	0.9019	0.8999	3.87
CaraNet (Lou et al. 2022)	0.8707	0.7958	0.9203	0.8621	0.8935	4.11
SSFormer-S (Wang et al. 2022)	0.8994	0.8363	0.9194	0.9086	0.9076	3.88
SSFormer-L (Wang et al. 2022)	0.9060	0.8500	0.9213	**0.9199**	0.9107	3.77
UNeXt (Valanarasu and Patel 2022)	0.7318	0.6284	0.7840	0.7656	0.7507	5.18
LDNet (Zhang et al. 2022)	0.8881	0.8208	0.9063	0.9046	0.8946	4.09
TGANet (Tomar et al. 2022)	0.8982	0.8330	0.9132	0.9123	0.9029	3.96
PVT-Cascade (Rahman and Marculescu 2023)	0.8950	0.8329	0.9355	0.8888	0.9103	3.90
PrototypeLab	**0.9086**	**0.8544**	0.9344	0.9136	**0.9194**	**3.71**

dataset was collected from two medical centers in Vietnam. It contains images captured using WLI, FICE, BLI and LCI. To train all the algorithms, we used 800 images for training, 100 images for validation, and 100 images for testing. Similarly, Kvasir-SEG consists of 1000 images collected from four hospitals in Norway. We have used 880 in the training set and rest 120 images in the validation and test set. Moreover, we use PolypGen (Ali et al. 2023) dataset collected from six medical centers in *Norway, United Kingdom, France, Italy and Egypt* as OOD dataset. PolypGen is collected from multiple hospitals that cover different clinical patient populations and modalities, which makes it diverse and useful for generalizability tests.

Experimental Setup: We have trained all the models on NVIDIA GeForce RTX 3090 GPU. All the images are first resized to 256×256 pixels. The training images are followed by simple data augmentation strategies, which includes random rotation, vertical flipping, horizontal flipping, and coarse dropout, are used to improve generalization and prevent overfitting. All models are trained on a similar hyperparameters configuration with a learning rate of $1e^{-4}$, batch size of 16, and an ADAM optimizer. We use a combination of binary cross-entropy and dice loss with equal weights as a loss function. In addition, we use an early stopping and *ReduceLROnPlateau* to avoid overfitting.

Results on BKAI-IGH dataset: Table 1 shows the results of all the models on BKAI-IGH dataset. The table demonstrates the superiority of PrototypeLab with a high DSC of 0.9243, mIoU of 0.8744, a high recall of 0.9194, a precision of 0.9494, low HD of 2.70. It outperforms 16 SOTA methods. DuAT (Tang et al. 2022) and SSFormer-L (Wang et al. 2022) are the most competitive network to our network, where our network still outperforms DuAT and SSFormer-L by 1.03% and 1.19% in DSC respectively. These results

Table 3: Result of models trained on BKAI-IGH (Lan et al. 2021) and tested on PolypGen (Ali et al. 2023).

Method	mDSC	mIoU	Recall	Precision	F2	HD
U-Net (Ronneberger, Fischer, and Brox 2015)	0.5841	0.5102	0.6142	0.7746	0.5739	4.36
DeepLabV3+ (Chen et al. 2018)	0.6757	0.6051	0.7074	0.8237	0.6732	4.03
PraNet (Fan et al. 2020)	0.7330	0.6659	0.7825	0.8182	0.7391	3.71
MSNet (Zhao, Zhang, and Lu 2021)	0.6777	0.6133	0.6811	**0.8881**	0.6657	4.06
TransFuse-S (Zhang, Liu, and Hu 2021)	0.6510	0.5720	0.6894	0.7952	0.6416	4.04
TransFuse-L (Zhang, Liu, and Hu 2021)	0.6592	0.5881	0.6792	0.8289	0.6487	4.10
Polyp-PVT (Dong et al. 2021)	0.7421	0.6746	0.7717	0.8494	0.7358	3.67
UACANet (Kim, Lee, and Kim 2021)	0.7063	0.6404	0.7265	0.8519	0.7016	3.87
DuAT (Tang et al. 2022)	0.7225	0.6553	0.7710	0.8081	0.7204	3.73
CaraNet (Lou et al. 2022)	0.6977	0.6329	0.7035	0.8830	0.6873	3.99
SSFormer-S (Wang et al. 2022)	0.7332	0.6664	0.7672	0.8386	0.7288	3.72
SSFormer-L (Wang et al. 2022)	0.7426	0.6732	0.7901	0.8209	0.7418	3.60
UNeXt (Valanarasu and Patel 2022)	0.3484	0.2669	0.4519	0.4834	0.3492	5.09
LDNet (Zhang et al. 2022)	0.6922	0.6193	0.7389	0.8013	0.6836	3.83
TGANet (Tomar et al. 2022)	0.6925	0.6206	0.7466	0.7833	0.6891	3.79
PVT-Cascade (Rahman and Marculescu 2023)	0.7271	0.6572	**0.8154**	0.7672	0.7358	**3.58**
PrototypeLab	**0.7583**	**0.6957**	0.7897	0.8456	**0.7563**	3.68

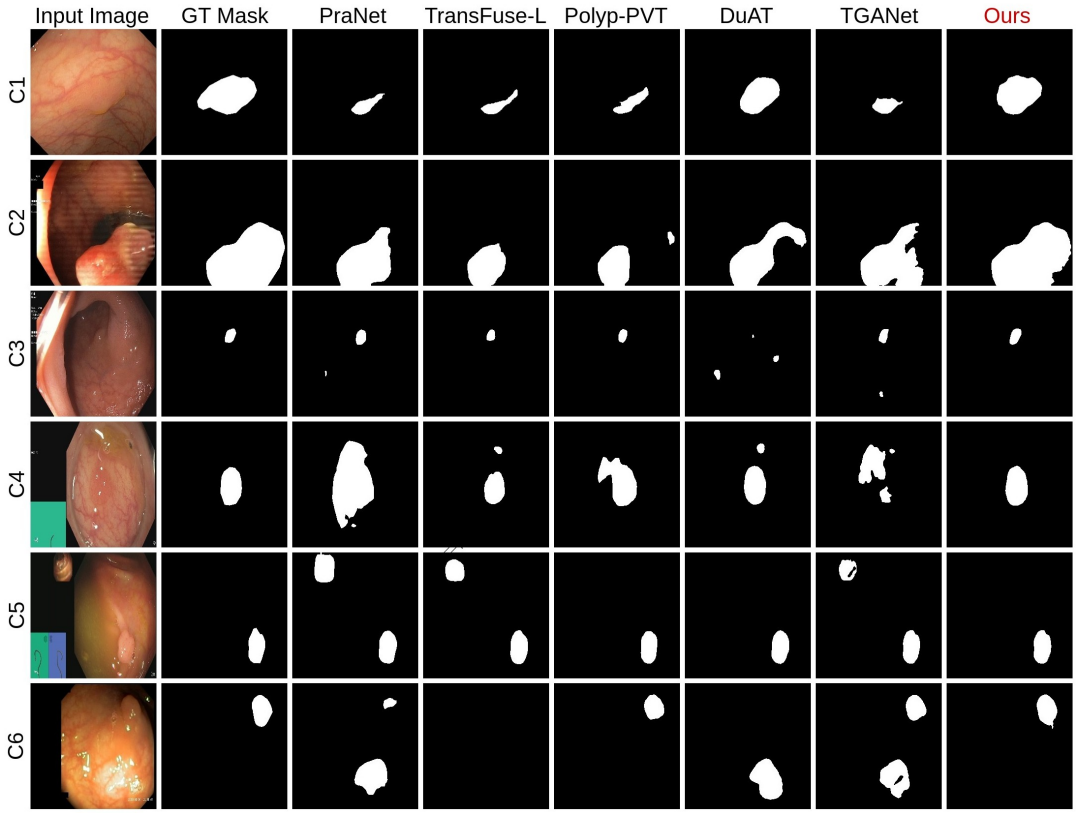

Figure 3: Qualitative results of models trained BKAI-IGH and tested on PolypGen. It can be observed that PrototypeLab produces a more accurate segmentation map in all the centers from C1 to C6.

suggest that PrototypeLab is highly effective in segmenting polyps on different endoscopic imaging techniques such as WLI, BLI, FCI, and FICE.

Results on Kvasir-SEG dataset: Table 2 shows the results of all the models on Kvasir-SEG dataset. PrototypeLab obtains a high DSC of 0.9086, mIoU of 0.8544, recall of 0.9344, precision of 0.9136, and low HD of 3.71. The most competitive network to PrototypeLab is SSFormer-L (Wang et al. 2022). Our model surpasses by SSFormer-L by 0.26% in DSC, 0.44% in mIoU, 0.06 in HD metrics. Thus, PrototypeLab surpasses all SOTA in both overlap based metrics and distance based metrics.

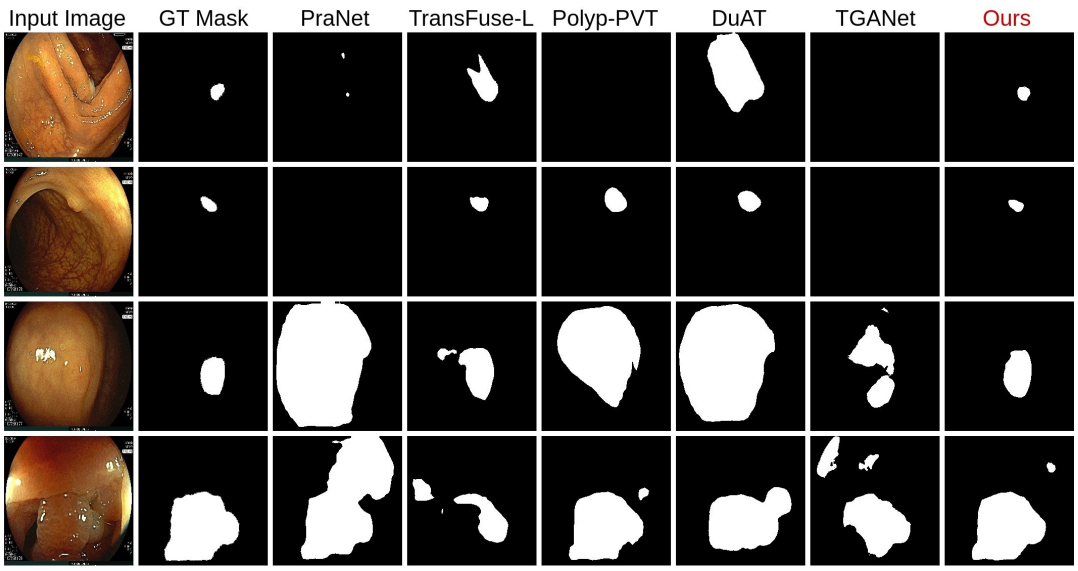

Figure 4: Qualitative results of the models trained on Kvasir-SEG and tested on BKAI-IGH.

Table 4: Ablation study of PrototypeLab on Kvasir-SEG. Here, BL = Baseline.

No	Method	mDSC	mIoU	Recall	HD
#1	BL (PVT-encoder + Decoder)	0.8971	0.8399	0.9203	3.88
#2	BL + CMGM	0.8971	0.8402	0.9199	3.84
#3	BL + (CMGM w/o LKDC) + PGM + PMGM	0.8935	0.8332	0.9282	4.02
#4	BL + CMGM + (PGM w/o EFFM) + PMGM	0.9012	0.8440	0.9284	3.82
#5	BL + (CMGM w/o LKDC) + (PGM w/o EFFM) + PMGM	0.9015	0.8464	0.9228	3.82
#6	BL + CMGM + PGM + PMGM (PrototypeLab)	**0.9086**	**0.8544**	**0.9344**	**3.71**

Table 5: The study provides model parameters, flops, and FPS for both SOTA methods and proposed PrototypeLab. 'Red', 'Green', and 'Blue' represent the best, second best and third best scores.

Method	Backbone	Param. (Million)	Flops (GMac)	FPS
U-Net	-	31.04	54.75	**94.99**
DeepLabv3+	ResNet50	39.76	43.31	60.47
PraNet	Res2Net50	32.55	6.93	29.68
MSNet	Res2Net50	29.74	8.98	38.12
TransFuse-S	ResNet34 + DeiT-S	26.35	11.5	34.76
TransFuse-L	ResNet50 + DeiT-B	143.74	82.71	37.6
Polyp-PVT	PVTv2-B2	25.11	5.30	47.54
UACANet	Res2Net50	69.16	31.51	26.04
DuAT	PVTv2-B2	24.97	**5.24**	44.56
CaraNet	Res2Net101	46.64	11.48	20.67
SSFormer-S	MiT-PLD-B2	29.57	10.1	49.11
SSFormer-L	MiT-PLD-B4	66.22	17.28	28.21
UNeXt	-	**1.47**	0.5695	88.01
LDNet	Res2Net50	33.38	33.14	31.14
TGANet	ResNet50	19.84	41.88	26.21
PVT-Cascade	PVTv2-B2	35.27	8.15	39.09
PrototypeLab	PVTv2-B2	51.34	174.87	23.85

Results of BKAI-IGH models tested on PolypGen: Table 3 shows results of the model trained on BKAI-IGH and tested on PolypGen (all centers combined). PrototypeLab obtains the highest DSC, mIoU, and F2 of 0.7583, 0.6957 and 0.7563, respectively. SSFormer-L and Polyp-PVT are the most competitive network, where PrototypeLab outperforms both networks by 1.57% and 1.62%, respectively in terms of DSC. Although PVT-Cascade has the least HD, our

Table 6: Results of models trained on Kvasir-SEG and tested on PolypGen. 'Red', 'Green', and 'Blue' represent the best, second best and third best scores.

Method	mDSC	mIoU	Recall	Precision	F2	HD
U-Net	0.5995	0.5347	0.6829	0.7523	0.6105	4.09
DeepLabV3+	0.7051	0.6389	0.8042	0.7659	0.7224	3.66
PraNet	0.7258	0.6632	0.7999	0.8217	0.7399	3.67
MSNet	0.7112	0.6500	0.7570	**0.8641**	0.7150	3.72
TransFuse-S	0.6956	0.6229	0.8079	0.7481	0.7180	3.66
TransFuse-L	0.7273	0.6621	0.8060	0.7822	0.7404	3.53
Polyp-PVT	0.7424	0.6763	0.8369	0.7985	0.7557	3.44
UACANet	0.7185	0.6517	0.7972	0.8007	0.7288	3.61
DuAT	0.7332	0.6654	0.8664	0.7342	0.7609	3.48
CaraNet	0.6976	0.6240	0.8264	0.7419	0.7285	3.68
SSFormer-S	0.7389	0.6734	0.8625	0.7496	0.7634	3.43
SSFormer-L	0.7556	0.6926	0.8530	0.7910	0.7703	3.37
UNeXt	0.4552	0.3761	0.6135	0.5600	0.4805	4.55
LDNet	0.7273	0.6604	0.8336	0.7685	0.7462	3.53
TGANet	0.7030	0.6386	0.8030	0.7654	0.7177	3.59
PVT-Cascade	0.7151	0.6460	0.8651	0.7108	0.7396	3.48
PrototypeLab	**0.7560**	**0.6966**	0.8603	0.7846	**0.7745**	**3.35**

Table 7: Ablation study of PrototypeLab on OOD dataset. The methods are trained on Kvasir-SEG and tested on PolypGen.

No	Method	mDSC	mIoU	HD
#1	Baseline (PVT-encoder + Decoder)	0.7503	0.6879	**3.36**
#2	Baseline + CMGM	**0.7585**	**0.6956**	**3.36**
#3	Baseline + (CMGM w/o LKDC) + PGM + PMGM	0.7234	0.6566	3.49
#4	Baseline + CMGM + (PGM w/o EFFM) + PMGM	**0.7538**	0.6909	**3.36**
#5	Baseline + (CMGM w/o LKDC) + (PGM w/o EFFM) + PMGM	0.7521	**0.6910**	**3.39**
#6	Baseline + CMGM + PGM + PMGM (PrototypeLab)	**0.7560**	**0.6966**	**3.35**

results is very competitive. We have similar findings when the model is trained on Kvasir-SEG and tested on PolypGen (Table 6), which suggests that PrototypeLab is more effective in handling OOD datasetst.

Figure 3 and 4 shows the qualitative results comparison of different models on diminutive polyp, flat polyp and regular polyp. The most competitive network, DuAT and Polyp-PVT exhibit over-segmentation or under-segmentation on different scenarios. However, PrototypeLab can segment more accurately on diminutive, flat and noisy images as compared to the SOTA baselines. Table 5 shows the number of parameters, flops and processing speed. The Table shows that PrototypeLab has 51.34 Million parameters and 174. 87 GMac Flops with a processing speed of 23.85. Although the processing speed is close to near real-time ($\approx 30 fps$), our architecture is more accurate, which is essential in clinical settings for early diagnosis and treatment. Therefore, the trade-off between speed and increased accuracy can be compensated.

Ablation study: Table 4 shows the ablation study of the PrototypeLab on the Kvasir-SEG. The results show that the baseline (#1) obtains DSC of 0.8971, mIoU of 0.8399, and HD of 3.88. In setting #2, a slight performance improvement was observed. This is because the masks generated by the CMGM were not used in the decoder, but were used by the PGM to generate multiple prototypes. These prototypes were then utilized by the PMGM to generate multiple prototype mask in setting #6, which explains the limited performance improvement when comparing setting #2 to setting #1.To demonstrate the impact of LKDC block and EFFM, we have conducted three experiments in setting #3, #4 and #5, where we can observed a drop in performance when compared with setting #6. Specifically, when both the LKDC block and EFFM were removed in setting #5, a 0.71% decrease in DSC, a 0.80% decrease in mIoU, and a 0.11% increase in HD were observed compared to setting #6 The Table shows that the baseline is improved by adding CMGM and further improved by adding PGM + PMGM. PrototypeLab (#6) offers an improvement of 1.15% in DSC, 1.45% in mIoU and 0.17% in HD when compared with baseline.

Conclusion

We propose a prototype learning based new segmentation model, called PrototypeLab, for in-distribution and out-of-distribution polyp segmentation. The use of prototypes in the proposed architecture helps in dealing with variability present in the medical images making the model more robust to inter-patient variations. It helps to perform well on diminutive, flat, partially visible, noisy images and camouflage properties of polyp. The proposed architecture obtains high DSC of 0.9243, mIoU of 0.8744, and low HD of 2.70 on the BKAI-IGH dataset. Our extensive experiments revealed that PrototypeLab exhibits superior performance compared to 16 state-of-the-art methods across three distinct datasets, including notoriously difficult multi-center

out-of-distribution datasets. In the future, we aim to develop PrototypeLabV2, by further optimizing speed and accuracy for mobile applications.

References

Ali, S.; Jha, D.; Ghatwary, N.; Realdon, S.; Cannizzaro, R.; Salem, O. E.; Lamarque, D.; Daul, C.; Riegler, M. A.; Anonsen, K. V.; et al. 2023. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1): 75.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Corley, D. A.; Jensen, C. D.; Marks, A. R.; Zhao, W. K.; Lee, J. K.; Doubeni, C. A.; Zauber, A. G.; de Boer, J.; Fireman, B. H.; Schottinger, J. E.; et al. 2014. Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine*, 370(14): 1298–1306.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *Proceedings of the International conference on medical image computing and computer-assisted intervention (MICCAI)*, 263–273.

Jha, D.; Riegler, M. A.; Johansen, D.; Halvorsen, P.; and Johansen, H. D. 2020a. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *Proceedings of the International symposium on computer-based medical systems (CBMS)*, 558–564.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; Lange, T. d.; Johansen, D.; and Johansen, H. D. 2020b. Kvasir-SEG: A segmented polyp dataset. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, 451–462.

Kim, T.; Lee, H.; and Kim, D. 2021. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the ACM International Conference on Multimedia*, 2167–2175.

Lan, P. N.; An, N. S.; Hang, D. V.; Van Long, D.; Trung, T. Q.; Thuy, N. T.; and Sang, D. V. 2021. NeoUNet: Towards accurate colon polyp segmentation and neoplasm detection. *arXiv preprint arXiv:2107.05023*.

Lou, A.; Guan, S.; Ko, H.; and Loew, M. H. 2022. CaraNet: Context Axial Reverse Attention Network for Segmentation of Small Medical Objects. In *Proceedings of the Medical Imaging 2022: Image Processing*, volume 12032, 81–92.

Ng, K.; P. May, F.; and Schrag, D. 2021. US Preventive Services Task Force Recommendations for Colorectal Cancer Screening. *Journal of American Medical Association*, 325(19): 1943–1945.

Rahman, M. M.; and Marculescu, R. 2023. Medical Image Segmentation via Cascaded Attention Decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6222–6231.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 234–241.

Sanderson, E.; and Matuszewski, B. J. 2022. FCN-transformer feature fusion for polyp segmentation. In *Proceedings of the Annual Conference on Medical Image Understanding and Analysis (MIUA 2022)*, 892–907.

Siegel, R. L.; Miller, K. D.; Wagle, N. S.; and Jemal, A. 2023. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1): 17–48.

Tang, F.; Huang, Q.; Wang, J.; Hou, X.; Su, J.; and Liu, J. 2022. DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation. *arXiv preprint arXiv:2212.11677*.

Tomar, N. K.; Jha, D.; Bagci, U.; and Ali, S. 2022. TGANet: Text-guided attention for improved polyp segmentation. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, 151–160.

Valanarasu, J. M. J.; and Patel, V. M. 2022. Unext: Mlp-based rapid medical image segmentation network. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, 23–33.

Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; and Song, S. 2022. Stepwise feature fusion: Local guides global. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, 110–120.

Zhang, R.; Lai, P.; Wan, X.; Fan, D.-J.; Gao, F.; Wu, X.-J.; and Li, G. 2022. Lesion-Aware Dynamic Kernel for Polyp Segmentation. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, 99–109.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 14–24.

Zhao, X.; Zhang, L.; and Lu, H. 2021. Automatic polyp segmentation via multi-scale subtraction network. In *Proceedings of the International Conference Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 120–130.