

XP-Net: An Attention Segmentation Network by Dual Teacher Hierarchical Knowledge distillation for Polyp Generalization

Ragu B¹, Antony Raj^{1,2}, Rahul GS^{1,2}, Sneha Chand^{1,2}, Preejith SP¹ and Mohanasankar Sivaprakasam²

¹Healthcare Technology Innovation Centre, Chennai, India

²Department of Electrical Engineering, IIT Madras, Chennai, India

Abstract

-5pt Endoscopic imaging is largely used as the diagnostic tool for Colon polyps-induced GI tract cancer. This diagnosis via image identification requires expertise that may be lacking in inexperienced physicians. Hence, using a software-aided approach to detect those anomalies may better identify the tissue abnormalities. In this paper, a novel deep learning network 'XP-Net' with Effective Pyramid Squeeze Attention (EPSA) module using hierarchical adversarial knowledge distillation by a combination of two teacher networks is proposed. It adds 'complementary knowledge' to the student network- thus aiding in the improvement of network performance. The lightweight EPSA block enhances the current network architecture by capturing multi-scale spatial information of objects at a granular level with long-range channel dependency. The XP-Net compiled into the NVIDIA TensorRT engine gave a better real-time performance in terms of throughput. The proposed network has achieved a dice score of 0.839 and IoU of 0.805 in the validation data set, and it was able to attain an average throughput of 60 fps in mobile GPU. This proposed deep learning-based segmentation approach is expected to aid clinicians in addressing the complications involved in the identification and removal of precancerous anomalies more competently.

Keywords

Polyp, Generalization, Attention block, Knowledge distillation

1. Introduction

Colorectal polyps are one of the early indicators of lower Gastro-Intestinal (GI) tract cancer. These polyps are extra growth lumps of tissues, having no particular function in the bodily processes [1]. Although these growth tissues are often benign, they can become cancerous. The early detection and removal of the polyps in the colon region may prevent these tissues from becoming cancerous. Colonoscopy is a general diagnostic procedure widely used to investigate the colon region for any type of malformation and disease [2]. Generally, a trained physician visually inspects the colon region for polyps and removes them using a minimally invasive endoscopic surgery. Research on the visual inspection of the colon region shows that small size adenomas (benign tumor), less than 5mm diameter, have a miss rate of 27% and for adenomas greater than 10mm have a miss rate of 6%. It has been reported that the quality of bowel preparation and the experience of colonoscopists are major contributory factors to missed polyps during a colonoscopy [3]. A quick alternative, computer-vision based polyps de-

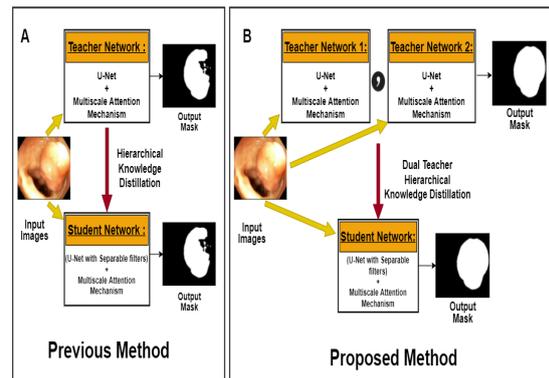


Figure 1: (A) shows a generic hierarchical knowledge distillation using a single teacher and (B) is our proposed methodology using dual teacher to derive the student network

tection is a highly researched area that has been found effective to mitigate the miss rates and assist in the faster diagnosis for colonoscopists [4]. The addition of deep learning techniques proves to be much more effective, since a network like U-net has shown promising results in biomedical imaging and widely accepted as the state-of-the-art image to image translation network [5].

In this paper, the U-net was chosen as the baseline model because of its ability to outperform other segmentation networks with extensive data augmentation re-

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

ragu.b@htic.iitm.ac.in (R. B); antony.raj@htic.iitm.ac.in (A. Raj)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)



ardless of a limited dataset, as reported Ronneberger et al. [5]. A plug-and-play EPSA module was implemented as proposed by EPSANet [6] with U-Net for enhancing the multiscale spatial information, which results in the detection of objects over different scale factors [6]. Since the baseline U-Net with the EPSA module was found to be computationally heavy for real-time performance, model compression techniques through knowledge distillation are implemented [7]. Among the other model compression approaches, knowledge distillation shows great superiority, which is to transfer knowledge of a large teacher model to a small student model [8]. In our proposed student network, we implemented separable filters resulting in model reduction by 78% of the teacher network. We implemented a hierarchical knowledge distillation technique which was proposed in the paper HAD-Net [8] where a single teacher network is used to distill the knowledge. Whereas, in our proposed methodology, the Dual teacher transfers the complementary knowledge to the student network. All the models were trained over EndoCV2022 Challenge dataset [9][10][11].

In summary, the contributions in this paper are as follows:

- A hierarchical dual teacher knowledge distillation network to transfer the complementary knowledge of both networks to a student.
- A student network with a lower computational cost for real-time performance without significantly reducing accuracy.
- Experiments: By evaluating our model's generality in the external Kvasir-Seg dataset [12], The dice and IoU scores of 0.782 and 0.769 are achieved, respectively.

2. XP-Net

2.1. Methodology

In our methodology, a student network is derived from two teacher networks through a hierarchical knowledge distillation process. The two teacher networks that are highly computational, transfer their complementary knowledge to a lightweight student network. The baseline U-Net architecture has the ability to capture features at multiple scales. To enhance this visual perception of the U-Net network, we implemented an Effective Pyramidal Squeeze Attention (EPSA) block at the first encoder of the U-Net. This attention mechanism boosts the allocation of the most informative feature expressions while suppressing the less useful ones, allowing the model to focus on clinically crucial areas [6]. The lightweight EPSA block enhanced the current architecture's ability by capturing multi-scale spatial information of objects at

a granular level with long-range channel dependency at the initial stage of the network.

Our first teacher network comprises of U-Net with EPSA module. Similarly, we trained the second teacher network, a baseline U-Net with EPSA block using pix2pix GAN [13], which has a promising result for an image to image translation that learns a loss adapted to the input data and task. The proposed student network consists of separable filters that hold the same U-Net architecture with the EPSA module, which results in the reduced number of learnable parameters from the defined teacher network. The hierarchical knowledge distillation technique used in our method is proposed in [8], where a single teacher is used for knowledge distillation. However the network that we have developed utilizes the dual teachers via multi-step learning as suggested in [14] to map the in-between features to train the respective student network.

The input and target of the teacher and student network is denoted by x and y . The output segmentation of two teachers and student is denoted by $T_{(1,2)}\hat{y}$ and $S\hat{y}$ respectively. The multi-scale feature map of teacher and student is denoted by $T_{(1,2)}y_{latent}$ and $S\hat{y}_{latent}$. In hierarchical knowledge distillation, the student loss is denoted by L_s which consists of weighted combination of two terms, (a) the sum of dice [15] and tversky loss [16] with the student generated segmentation ($S\hat{y}$) and ground truth (y), (b) mean square error adversarial loss. The overall student loss is given in equation 2.

$$DV = [\text{Dice Loss} + \text{Tversky Loss}] \quad (1)$$

$$L_s = DV[S\hat{y}; y] + \lambda * MSE[HD(x, S\hat{y}, S\hat{y}_{latent}), 1] \quad (2)$$

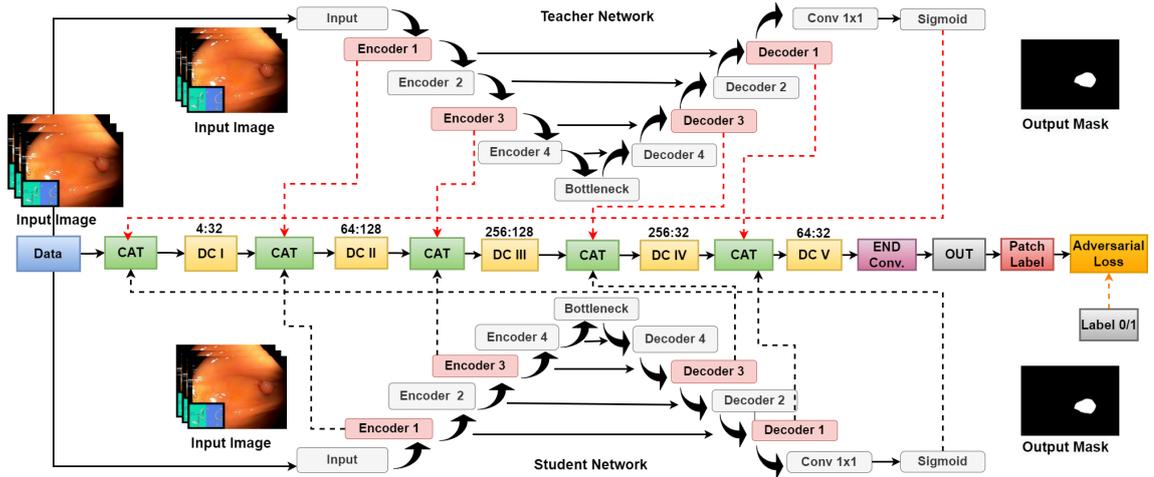
The hierarchical discriminator (HD) is trained using LS-GAN loss denoted as L_{HD} . The L_{HD} is made up of two mean square error term. one term is between the HD output after being passed a "fake" datasample from the teacher, and a tensor of all zeros [17]. The other term is a mean square error loss between the HD output after being passed a "real" data sample from either teacher 1 or teacher 2 and a tensor of all ones. The overall discriminator loss is denoted in equation 3.

$$L_{HD} = MSE[HD(x, S\hat{y}, S\hat{y}_{latent}), 0] + MSE[HD(X, T_{(1,2)}\hat{y}, T_{(1,2)}\hat{y}_{latent}), 1] \quad (3)$$

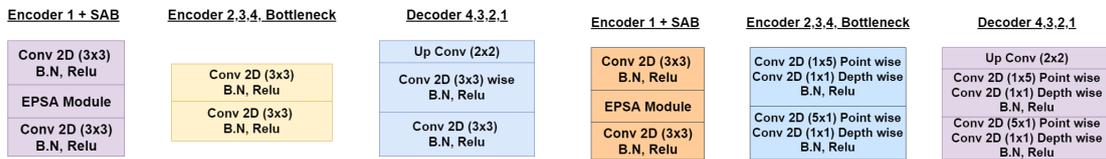
2.2. Network Architecture

2.2.1. Teacher Network

CABR32-P-CBR64-P-CBR128-P-CBR256-P-CBR512-UPCONV256-CBR256-UPCONV128-



(a) Teacher and Student Discriminator Network



CABR32-P-CBR64-P-CBR128-P-CBR256-P-CBR512
 -UPCONV256-CBR256-UPCONV128-CBR128
 -UPCONV64-CBR64-UPCONV32-CBR32-CE1

(b) Teacher Network Blocks

CAPDBR32-P-CPDBR64-P-CPDBR128-P-CPDBR256
 -P-CPDBR512-UPCONV256-CPDBR256-UPCONV128-CPDBR128
 -UPCONV64-CPDBR64-UPCONV32-CPDBR32-CE1

(c) Student Network Blocks

Figure 2: XP-Net Network Architecture

CBR128-UPCONV64-CBR64-UPCONV32-CBR32-CE1

- CABRK

CABRK represents two stacks of Convolution, Batch norm and Relu activation function with K number of output filters with an intermediate attention block

- CBRK

CBRK represents two stacks of Convolution, Batch norm and Relu activation function with K number of output filters.

- CEK

CEK denotes a (1,1) convolution with k output feature map with a Sigmoid activation function.

- UPCONVK

UPCONVK represents a layer of transpose convolution with a kernel size (2,2), stride (2,2) with k output number of feature maps.

- Pool (P)

Pool represents a pooling layer with a kernel size (2,2) and stride (2,2).

2.2.2. Student Network

CAPDBR32-P-CPDBR64-P-CPDBR128-P-CPDBR256-P-CPDBR512-UPCONV256-CPDBR256-UPCONV128-CPDBR128-UPCONV64-CPDBR64-UPCONV32-CPDBR32-CE1

- CPDBRK

CPDBRK represents stack of (A) point wise convolution of kernel size (1,5) and depth wise convolution of kernel size (1,1) followed by Batch norm and Relu and (B) point wise convolution of kernel size (5,1) and depth wise convolution of kernel size (1,1) followed by Batch norm and Relu. All the convolution layers consists of K number of feature outputs.

- CAPDBRK

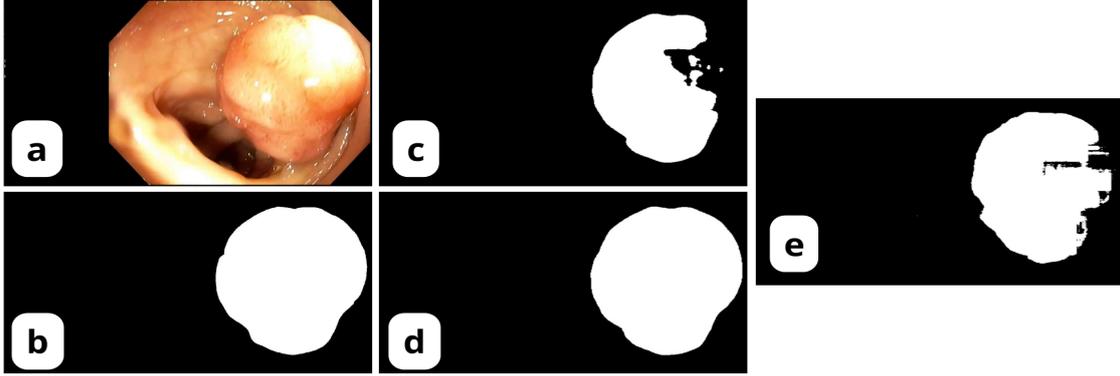


Figure 3: The input image (a) and ground truth (b) is given and student networks mask (e) which has learned from teacher 1 (c) and teacher 2 (d) network's.

CAPDBRK block is a modified version of CPDBRK block where attention block is placed in between the two sets of point wise convolution, depth wise convolution, batch norm and relu.

2.2.3. Discriminator Network

CAT-DC32-CAT-DC128-CAT-DC128-CAT-DC32-CAT-DC32-ENCONV

- CAT

CAT is the concatenation of two different layers either from teacher or student network.

- DCK

DCK represents a stack of convolution of kernel size (3,3), padding and stride of (1,1) with instance norm and Leaky Relu with negative slope of 0.2.

The hierarchical discriminator consists of five discriminator blocks (DC) and an End Convolution (ENCONV). In our proposed model, the feature map from encoder 1, encoder 3, decoder 1, decoder 3 from the teacher or student network are used for hierarchical knowledge distillation. The full network architecture is described in Fig.2.

3. Dataset and Implementation

3.1. Dataset

Automatic polyp detection and classification requires the availability of big datasets of polyp images or videos along with high-quality, manual annotations provided by experts. These annotations provide the ground truth necessary to train the supervised deep learning models.

Table 1

Comparison between teacher and student network

Dataset	Network	No. of params	Dice score	IoU score
EndoCV Dataset	Teach. 1	7,774,374	0.893	0.889
	Teach. 2	7,774,374	0.871	0.884
	Student	1,839,333	0.839	0.805
	U-Net	7,763,041	0.841	0.812
Kvasir Dataset	Teach. 1	7774374	0.812	0.809
	Teach. 2	7,774,374	0.803	0.798
	Student	1,839,333	0.783	0.769
	U-Net	7,763,041	0.798	0.784

EndoCV2022 challenge provided us with series of sequence dataset of 2631 images with their corresponding ground truth masks [9][10][11]. In that, we utilized more than 95% of the data for training and 5% of the data for testing. External dataset such as Kvasir-Seg was utilized for testing the model generality.

3.1.1. Dataset augmentation

All the models were trained with an input image size of 512x512. The data augmentation such as random rotate, horizontal flip, vertical flip, perspective transform was implemented. Usually the endoscopic images are subjected to different light sources that might have different intensities of brightness, contrast and hue, so images are augmented in such a way to replicate those scenarios.

3.2. Implementation

Both the teacher network is trained using Adam optimizer with initial learning rate of $3e^{-4}$ with step learning rate scheduler of gamma 0.1 and step size of 30. The networks were trained for 450 epoch with batch size of 8. The student network was trained using Adam opti-

mizer with β_1 0.5 and β_2 0.999 with a initial learning rate of $1e^{-4}$ with step learning rate scheduler of gamma 0.1 and step size of 30. After multiple experiments of initializing weights with uniform, xavier-uniform and kaiming-uniform given in pytorch weight initialization, it showed that kaiming uniform weight initialization have helped for better convergence of model. We also implemented our model in Nvidia TensorRT inference library for effective realtime model throughput. All the models were trained using Nvidia RTX 3090 GPU.

4. Results and Discussion

The networks were evaluated and the computed metrics are reported in Table.1. In the validation data of EndoCV dataset, the Teacher 1 model was able to achieve 0.893 and 0.889 for Dice score and IoU score, respectively. Similarly, the teacher 2 model was able to achieve 0.871 and 0.884 for the same metrics. The student network has achieved a commendable dice and IoU score of 0.839 and 0.805 even with the reduced number of learnable parameters. The trade-off here is the larger sized teacher network for a minimal loss in the accuracy of the light weight student network. Similarly, these metrics were calculated for Kvasir-Seg dataset and is reported in the Table 1.

Results have shown that the teacher 2 perform better for region with higher amount of specular reflection than teacher 1 for those regions. The student network thus obtains the complimentary knowledge from the two teacher networks. With reference to the ground truth, it is observed that the student network had proper segmentation even though one of the teachers had missed areas in its segmentation masks as shown in Fig 3. These results show that multiple teacher knowledge helps to generalize better segmentation.

As a part of benchmarking the network in terms of inference time, the model was converted into TensorRT engine for faster throughput. The model was able to attain an average throughput of 60 fps on GeForce RTX 3070 mobile GPU and 120 fps in Nvidia RTX 3090 GPU. From the results, we believe that constructing multiple teacher models which focuses on various aspects of the input data can distill a superior student network.

5. Conclusion

The proposed network is light weight and does faster computation when compared with traditional networks that are used for segmentation. Since this uses dual teachers for knowledge distillation, by increasing the number of teacher networks, there is room for further improvement in performance. Moreover, the sample size of data also plays a crucial role in the accuracy of the network. Further studies can be done to design a much more intelligent network for polyps and other varieties of early

cancer tissues.

References

- [1] B. Levin, D. A. Lieberman, B. McFarland, K. S. Andrews, D. Brooks, J. Bond, C. Dash, F. M. Giardiello, S. Glick, D. Johnson, et al., Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology, *Gastroenterology* 134 (2008) 1570–1595.
- [2] D. K. Rex, J. L. Petrini, T. H. Baron, A. Chak, J. Cohen, S. E. Deal, B. Hoffman, B. C. Jacobson, K. Mergener, B. T. Petersen, et al., Quality indicators for colonoscopy, *Gastrointestinal endoscopy* 63 (2006) S16–S28.
- [3] S. N. Bonnington, M. D. Rutter, Surveillance of colonic polyps: are we getting it right?, *World journal of gastroenterology* 22 (2016) 1925.
- [4] Y. Mintz, R. Brodie, Introduction to artificial intelligence in medicine, *Minimally Invasive Therapy & Allied Technologies* 28 (2019) 73–81.
- [5] 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.
- [6] H. Zhang, K. Zu, J. Lu, Y. Zou, D. Meng, Epsanet: An efficient pyramid squeeze attention block on convolutional neural network, *arXiv preprint arXiv:2105.14447* (2021).
- [7] G. Hinton, O. Vinyals, J. Dean, et al., Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* 2 (2015).
- [8] S. Vadalacchino, R. Mehta, N. M. Sepahvand, B. Nichyporuk, J. J. Clark, T. Arbel, Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images, in: *Medical Imaging with Deep Learning*, PMLR, 2021, pp. 787–801.
- [9] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N. T. Nguyen, D. Q. Tran, L. D. Huynh, N. Boutry, S. Rezvy, H. Chen, Y. H. Choi, A. Subramanian, V. Balasubramanian, X. W. Gao, H. Hu, Y. Liao, D. Stoyanov, C. Daul, S. Realdon, R. Cannizzaro, D. Lamarque, T. Tran-Nguyen, A. Bailey, B. Braden, J. E. East, J. Rittscher, Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, *Medical Image Analysis* 70 (2021) 102002.

URL: <https://doi.org/10.10162/j.media.2021.102002>.
doi:10.1016/j.media.2021.102002.

- [10] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.
- [11] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:10.48550/arXiv.2202.12031.
- [12] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, et al., Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, *Scientific data* 7 (2020) 1–14.
- [13] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [14] Z.-Q. Zhao, Y. Gao, Y. Ge, W. Tian, Orderly dual-teacher knowledge distillation for lightweight human pose estimation, arXiv preprint arXiv:2104.10414 (2021).
- [15] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, J. Li, Dice loss for data-imbalanced nlp tasks, arXiv preprint arXiv:1911.02855 (2019).
- [16] N. Nasalwai, N. S. Pun, S. K. Sonbhadra, S. Agarwal, Addressing the class imbalance problem in medical image segmentation via accelerated tversky loss function, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2021, pp. 390–402.
- [17] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.