Improving Synthetic Sample Selections with Method Inspired by Dynamic Programming

Jiajian Ma
Student ID: 222041049
School of Data Science
The Chinese University of Hong Kong, Shenzhen
222041049@link.cuhk.edu.cn

Abstract

Data augmentation is a proven method for addressing model generalization issues. However, most prior research has primarily focused on improving the generation of more realistic synthetic data, often overlooking a critical aspect: the selection of appropriate samples from synthetic data for subsequent model training. Traditional methods, such as random and prediction-based sampling, typically fail to consider the comprehensive distribution of both synthetic and real data, thereby leading to either limited diversity enhancement or biased data distribution. To address this issue, inspired by dynamic programming, we introduce a novel sampling strategy that integrates the distributions of both real and synthetic data to optimize training effectiveness. Experimentally, our strategy requires only a modest increase in polynomial time computational cost and significantly outperforms traditional methods. It improves model generalization on polyp segmentation tasks by 0.6% and 1.6% in mean Dice coefficient over random and prediction-based methods, respectively. Our code is available at https://github.com/497662892/Dynamic-Programming-Project.

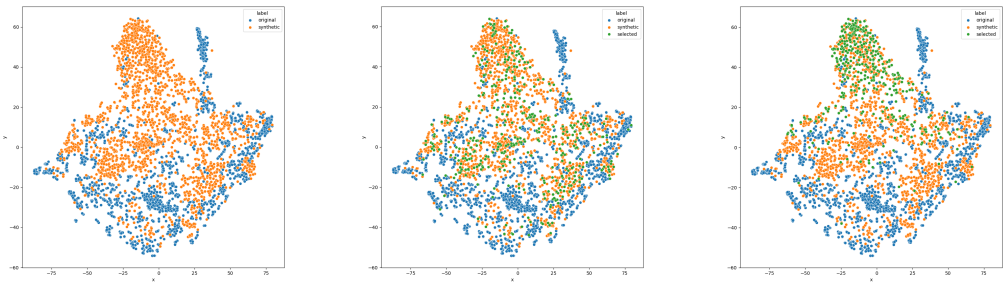
1 Introduction

Polyp segmentation models can effectively help doctor to identify polyps during endoscopic examinations, which is crucial for colon cancer screening [5]. However, these models are often trained on small scale labeled data from limited sources, constrained by data privacy and high costs of annotation [5]. These datasets frequently fail to capture the diversity of real-world clinical environments [1, 15, 7], leading to generalization problems in clinical application.

Previous research has proposed various methods to address the problem of model generalization[16, 4, 9, 8]. Among these, data augmentation-based methods have been proven effective at enhancing model generalization capabilities[16, 4, 8]. However, most prior studies have focused on improving the quality of synthetic images through model enhancements[3, 2, 12], overlooking a crucial aspect: selecting appropriate samples from the generated synthetic data for subsequent model training. Considering that the fundamental goal of data augmentation is to enrich the diversity of feature space distributions[16], the selection of suitable samples is critical to achieving enhanced generalization in augmented models.

For this issue, previous studies have commonly adopted two straightforward strategies for selecting synthetic samples. The first, a random selection strategy[12, 3, 2, 17], involves indiscriminately choosing samples from all generated data to include in the training set. The second, a prediction-based selection strategy[13, 8], involves selecting the worst-performing samples based on the current model's predictive outcomes to train the model further. However, both approaches have significant limitations.

Preprint. Under review.



(a) Data distribution (b) Random selection (c) Prediction based selection

Figure 1: The t-SNE visualization results of real and synthetic data in the feature space. (a) shows the distribution of the original data and synthetic data; (b) illustrates the results of the random sampling method, where the selected samples are relatively insufficient in out-of-distribution (OOD) areas; (c) shows the results of the prediction-based sampling method, where the selected samples are concentrated in OOD areas, leading to potential distribution bias.

With the random selection strategy, the choice of samples considers only the distribution of synthetic data, potentially overlooking areas of the real dataset that are out-of-distribution (OOD). This can lead to only a limited increase in feature space diversity, thereby inadequately enhancing the model's generalization performance on real-world data (Figure 1b). On the other hand, the prediction-based selection strategy tends to concentrate on samples from OOD areas, where the current model performs poorly. This method, while focusing on the real data distribution, may neglect the diversity of already selected synthetic samples. As a result, it could lead to bias in the data distribution, as it disproportionately adds samples from specific areas of the feature space (Figure 1c).

To address these limitations, inspired by dynamic programming, we propose a novel sampling strategy that considers both the distribution of real data and the distribution of previously selected synthetic data. Utilizing a greedy algorithm, our method selects, at each step, the point whose sum of average distances to real samples and to already selected synthetic samples in feature space is maximized. This approach achieves an efficient approximation for sample selection. Experimental results demonstrate that our sampling strategy significantly enhances the generalization performance of polyp segmentation models, surpassing traditional random sampling and prediction-based selection strategies, without incurring substantial additional computational costs.

To overcome these limitations, we have developed a novel sampling strategy inspired by dynamic programming, which consider both the distribution of real data and previously selected synthetic data. By employing a greedy algorithm, our method selects the point that maximize the sum of average distances to both real samples and previously selected synthetic samples in each step, thus achieving an approximation in polynomial time. Experimental results demonstrate that our sampling strategy significantly enhances the generalization performance of polyp segmentation models, surpassing traditional random sampling and prediction-based selection strategies, without incurring substantial additional computational costs.

2 Method

2.1 Preliminary

2.1.1 Polyp segmentation

Polyp segmentation is a classic task in medical image analysis, with the primary goal of automatically identifying the polyps in endoscopic images to prevent missed diagnoses during endoscopic examinations[5] (Figure2). In this project, we utilize the Dice Coefficient as the performance metric for our segmentation model[1]. The Dice Coefficient measures the similarity between two samples, making it ideal for evaluating binary classifications like polyp segmentation. It is defined as:

$$\text{Dice} = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

where X is the set of pixels predicted by the segmentation model and Y is the ground truth. A higher Dice score indicates better performance of the segmentation model.

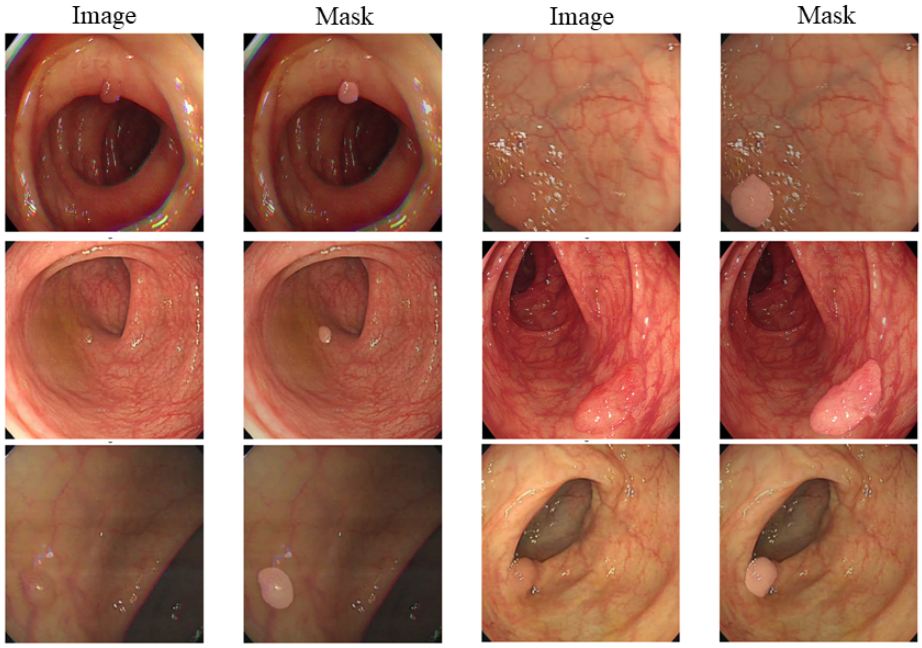


Figure 2: Examples of polyp segmentation.

2.1.2 Random sampling strategy

The following algorithm 1 describes the random sampling strategy which involves selecting a specific number of cases from the synthetic samples S_{syn} to form S_{select}. These selected samples S_{select} are then merged with the original dataset S_{ori} to create a new training set S_{train}, which is subsequently used to train the model. It is crucial to recognize that the distribution of the samples selected through this method strictly depends on the inherent distribution of the synthetic data.

Algorithm 1 Random Sampling Strategy for Data Augmentation

Require: Number of samples n, synthetic dataset S_{syn}, original dataset S_{ori}
Ensure: New training dataset S_{train}
$S_{\text{train}} \leftarrow S_{\text{ori}}$ ▷ Initialize S_{train} with S_{ori}
$S_{\text{select}} \leftarrow$ Randomly select n samples from S_{syn}
$S_{\text{train}} \leftarrow S_{\text{train}} \cup S_{\text{select}}$ ▷ Merge selected samples into S_{train}
return S_{train}

2.1.3 Prediction-based sampling strategy

The following algorithm 2 describes the prediction based sampling strategy, which leverages a model f_θ trained on the original dataset S_{ori}. This method predicts the performance of samples from the synthetic dataset S_{syn} and ranks them based on a specified metric. Samples with poorer predicted performance are prioritized for inclusion in the augmented training dataset S_{train}.

This strategy aims to enhance the training dataset by including samples that the model f_θ finds challenging, primarily including OOD samples related to the original distribution.

Algorithm 2 Prediction-Based Sampling Strategy for Data Augmentation

Require: Number of samples n, synthetic dataset S_{syn}, original dataset S_{ori}, trained model f_θ, performance metric metric
Ensure: New training dataset S_{train}
$S_{\text{train}} \leftarrow S_{\text{ori}}$ ▷ Initialize S_{train} with S_{ori}
scores $\leftarrow$ Evaluate f_θ on S_{syn} using metric ▷ Predict and score samples
$S_{\text{select}} \leftarrow$ Select n samples from S_{syn} with the worst scores based on metric
$S_{\text{train}} \leftarrow S_{\text{train}} \cup S_{\text{select}}$ ▷ Merge selected samples into S_{train}
return S_{train}

2.2 Dynamic programming formulation of our method

In our proposed sampling strategy, we need to consider the distribution of both real samples and selected synthetic samples. Therefore, we can formulate the problem as a sequential decision-making and dynamic programming problem, as shown below:

- **Background**:
  - Original Sample Set S_{ori}: Set of real samples.
  - Synthetic Sample Set S_{syn}: Set of synthetic samples.
- **State Definition**:
  - $X_n = (S_{\text{select}}, S_{\text{remain}})$
  - S_{select}: The set of selected synthetic data.
  - S_{remain}: The set of unselected synthetic data.
  - Initial state: S_{select} is $\emptyset$, $S_{\text{remain}} = S_{\text{syn}}$.
- **Decision/Action**:
  - Select a point a from S_{remain} and add it to S_{select}.
- **State Transition Function**:

$$X_{n+1} = f(X_n = (S_{\text{select}}, S_{\text{remain}}), A_n = a) = (S_{\text{select}} \cup \{a\}, S_{\text{remain}} \setminus \{a\})$$

- **Reward Function**:

$$g_n(X_n = (S_{\text{select}}, S_{\text{remain}}), A_n = a) = \frac{1}{|S_{\text{ori}}|} \sum_{s \in S_{\text{ori}}} \text{dist}(a, s) + \frac{1}{|S_{\text{select}}|} \sum_{s \in S_{\text{select}}} \text{dist}(a, s)$$

  Here, dist measures the distance between two points.

- **Value Function**:

$$\text{Sum of all rewards for actions taken: } V(n) = \sum_{i=1}^{n} g_i(X_i, A_i)$$

- **Termination Condition**:
  - When the sampling number limit is reached $n = \max_{\text{sampling}}$.
- **Goal**:
  - Maximize the value function for selecting samples from S_{syn}, ensuring that S_{select} can represents S_{ori} and maintains internal diversity.

Since images are high-dimensional data, for computational convenience, we project them into a 512-dimensional feature space using the encoder of a segmentation model trained on S_{ori}, and we adopt the L2-norm as the metric for the distance between different samples.

Algorithm 3 Our Synthetic Data Sampling Algorithm

Input:
S_{ori}: the set of original data
S_{syn}: the set of synthetic data
max_{sample}: maximum number of samples
$Dist(a, b)$: function to compute the distance matrix
Prepare:
$D_{\text{syn\_to\_ori}} = Dist(S_{\text{syn}}, S_{\text{ori}})$ ▷ distance matrix from synthetic to original
$D_{\text{syn\_to\_syn}} = Dist(S_{\text{syn}}, S_{\text{syn}})$ ▷ distance matrix from synthetic to synthetic
Initialize selected set: $S_{\text{select}} = \emptyset$
Initial Step:
S_{select}.add($\arg\max_{s \in S_{\text{syn}}} \{D_{\text{syn\_to\_ori}}[s].\text{mean}()\}$) ▷ maximize distance to S_{ori}
Loop:
while len(S_{select}) $< max_{\text{sample}}$ **do**
    $S_{\text{remain}} = S_{\text{syn}} \setminus S_{\text{select}}$
    $D_{\text{re\_to\_ori}} = D_{\text{syn\_to\_ori}}[S_{\text{remain}}, :]$ ▷ indexing with S_{remain}
    $D_{\text{re\_to\_sele}} = D_{\text{syn\_to\_syn}}[S_{\text{remain}}, S_{\text{select}}]$ ▷ indexing with S_{remain}, S_{select}
    S_{selected}.add($\arg\max_{s \in S_{\text{remain}}} \{D_{\text{re\_to\_ori}}[s].\text{mean}() + D_{\text{re\_to\_sele}}[s].\text{mean}()\}$)
end while
return S_{selected}

2.3 Approximation with greedy algorithm

Due to the exponential growth of the state space with the inclusion of more synthetic samples, we adopt a greedy algorithm to mitigate the computational burden. At each step of sample selection, the synthetic sample that maximizes the sum of its average distance to the real samples and its average distance to the already selected synthetic samples is chosen. The pseudocode of our algorithm is shown below:

The computational complexity per step is approximately $O(N \times T)$, where N is the total number of synthetic samples and T is the current number of selected synthetic samples $|S_{\text{select}}|$. The overall computational complexity does not exceed $O(N^3)$.

3 Experiment Setup

3.1 Polyp segmentation datasets

Our study utilized four publicly available datasets for polyp segmentation: SUN-SEG [11, 7] (49,136 images), Kvasir-SEG [6] (1,000 images), CVC-ClinicDB [14] (612 images), and CVC-ColonDB [14] (300 images). To expedite the validation of our work's efficacy, we sampled every 10th frame from each video in the SUN-SEG dataset and maintained the original partitioning scheme to derive the training set (1,993 images), validation set (1,442 images), and test set (1,574 images).

Among these, the SUN-SEG training set is used as S_{ori} for model training, the SUN-SEG validation set is used for model selection, and the remaining datasets are used as the testing set to evaluate the generalization performance of the trained model.

3.2 Data augmentation method

We employed a Stable Diffusion based inpainting method as the data augmentation strategy in this project, which can realistically inpaint polyps across different negative backgrounds and provide high-quality pseudo-masks[1].

We additionally used negative images from SUN-SEG (781 images) and LD-PolypVideo [10] (1,220 images) as backgrounds for inpainting. The method and number for generating synthetic data were

[1] https://github.com/497662892/Ai-Security-Final-Project

consistent with our previous work[1], resulting in a total of 1,993 synthetic images, of which 1,568 are qualified cases, forming our synthetic dataset S_{syn}.

In this project, we set the max number of sampling for synthetic data to 400 images. We selected S_{select} through random sampling strategy, prediction-based sampling strategy, and our proposed sampling method, and merged them with the real dataset S_{ori} (i.e., the SUN-SEG training set) to obtain three different training sets S_{train}.

3.3 Segmentation model

We adapted the polyp-PVT model, a robust polyp segmentation method [1], as our segmentation model. The baseline model and the augmented model were trained on the SUN-SEG training set S_{ori} and the combined training set S_{train}, respectively. All models were trained following the protocols proposed by Dong et al. [1]. The model demonstrating the highest performance during validation was advanced to the testing phase.

To evaluate the models' generalization performance, we tested them across four datasets, including one internal (SUN-SEG test set) and four external test sets (Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB). The Dice coefficient was employed as the metric to assess the segmentation model's performance.

4 Result

4.1 Generalization in polyp segmentation

Table 1 displays the polyp-PVT's performance across diverse datasets with or without augmentation strategies using different sampling methods. Implementing our data augmentation method results in significant performance enhancements on both the SUN-SEG and external datasets. As indicated in the last column of Table 1, the model trained with data augmentation displayed a uniform improvement in the overall mDice: with +1.3% on random sampling (0.8448 vs. 0.8313), +0.3% on prediction-based sampling (0.8343 vs. 0.8313), and +1.7% on our proposed sampling method (0.8507 vs. 0.8313).

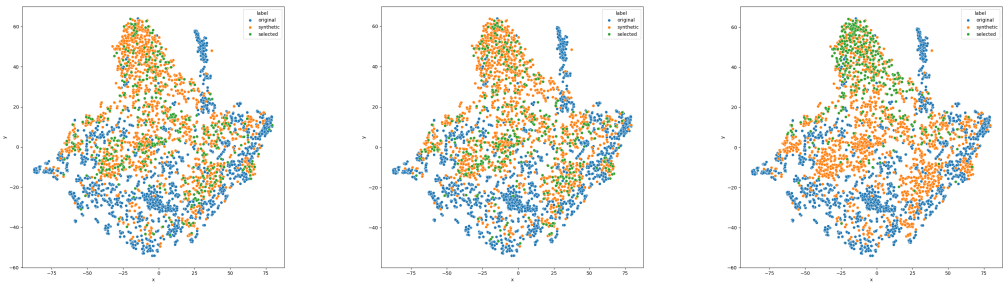
More importantly, our proposed sampling strategy not only outperformed random sampling (+0.6%, 0.8507 vs. 0.8448) and prediction-based sampling (+1.6%, 0.8507 vs. 0.8343) in terms of overall mDice, but also achieved the best performance across all external datasets (column 3-5). These findings affirm that our sampling method can considerably boost model generalization capabilities.

Table 1: Comparison of mDice Across Different Data Augmentation Methods

Dataset	SUN-SEG	Kvasir-SEG	CVC-ClinicDB	CVC-ColonDB	Overall
No Aug	0.8254	0.8143	0.8358	0.8498	0.8313
Random	**0.8450**	0.8351	0.8496	0.8495	0.8448
Prediction	0.8281	0.8125	0.8476	0.8490	0.8343
Ours	0.8360	**0.8457**	**0.8630**	**0.8579**	**0.8507**

4.2 Visualization of the sampling strategies

The sample distributions obtained using different synthetic sample selection strategies are shown in the Figure 3. Intuitively, we can observe that, compared to the random selection strategy, our proposed strategy can collect more samples in the OOD regions, thereby theoretically increasing the diversity of the feature space more effectively (Figures 3a and 3b. Compared to the prediction-based selection method, our strategy effectively avoids clustering of selected sample points and maintains the integrity of feature space distribution (Figure 3c and 3b).



(a) Random selection (b) Ours (c) Prediction based selection

Figure 3: The t-SNE visualization of different synthetic data sampling methods.

4.3 Sampling efficiency of our method

The time complexity of our proposed synthetic sample selection algorithm is approximately $O(N^3)$, where N is the number of synthetic samples. In our experiments, it took approximately 10 seconds to complete the full sorting of 1,993 sample points in a single CPU and Jupyter Notebook environment.

5 Conclusion

In this project, inspired by dynamic programming, we proposed a sampling strategy that considers both the distribution of real data and previously selected synthetic data. We demonstrated its computational feasibility and its superiority over traditional random selection and prediction-based sampling strategies in enhancing model generalization for polyp segmentation tasks.

However, our work has some limitations. First, our validation was limited to one data augmentation method, a single set of samples, one polyp segmentation model, and four datasets, suggesting that more extensive experiments are necessary to confirm the effectiveness of our approach. Second, our sampling process still requires $O(N^3)$ computational overhead, which may need further optimization for handling a large number of synthetic samples. Finally, the applicability of our approach to machine learning tasks beyond polyp segmentation remains to be explored.

References

[1] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[2] Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. Arsdm: Colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2023.

[3] Jan Andre Fagereng, Vajira Thambawita, Andrea M Storås, Sravanthi Parasa, Thomas De Lange, Pål Halvorsen, and Michael A Riegler. Polypconnect: Image inpainting for generating realistic gastrointestinal tract images with polyps. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 66–71. IEEE, 2022.

[4] Rutger HJ Fick, Alireza Moshayedi, Gauthier Roy, Jules Dedieu, Stéphanie Petit, and Saima Ben Hadj. Domain-specific cycle-gan augmentation improves domain generalizability for mitosis detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 40–47. Springer, 2021.

[5] Dexin Gong, Lianlian Wu, Jun Zhang, Ganggang Mu, Lei Shen, Jun Liu, Zhengqiang Wang, Wei Zhou, Ping An, Xu Huang, et al. Detection of colorectal adenomas with a real-time computer-aided system (endoangel): a randomised controlled study. *The lancet Gastroenterology & hepatology*, 5(4):352–361, 2020.

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020.

[7] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022.

[8] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024.

[9] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022.

[10] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 387–396. Springer, 2021.

[11] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4):960–967, 2021.

[12] Alexander K Pishva, Vajira Thambawita, Jim Torresen, and Steven A Hicks. Repolyp: A framework for generating realistic colon polyps with corresponding segmentation masks using diffusion models. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 47–52. IEEE, 2023.

[13] Zhenrong Shen, Xi Ouyang, Bin Xiao, Jie-Zhi Cheng, Dinggang Shen, and Qian Wang. Image synthesis with disentangled attributes for chest x-ray nodule augmentation and detection. *Medical Image Analysis*, 84:102708, 2023.

[14] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[15] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 699–708. Springer, 2021.

[16] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

[17] Lei Zhou. Spatially exclusive pasting: A general data augmentation for the polyp segmentation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 01–07. IEEE, 2023.