Distribution-Aware Synthetic Sample Selection for Enhanced Model Generalization in Polyp Segmentation

Jiajian Ma
Student ID: jm10850
Center for Data Science
New York University, New York
jm10850@nyu.edu

Abstract

Polyp segmentation models often struggle to generalize effectively across varying clinical environments, limiting their utility in real-world applications. Data augmentation offers a straightforward solution, but existing sampling strategies—typically random or prediction-based—fail to fully exploit the diversity of synthetic data. We propose a distribution-aware synthetic sample selection strategy that jointly considers real and synthetic data distributions, ensuring a balanced coverage of the feature space. By selecting synthetic samples that maximize their feature-space distance to both the original dataset and previously chosen synthetic samples, our method enhances model robustness without incurring excessive computational overhead. Experiments on polyp segmentation tasks show that our approach improves mean Dice coefficients by up to 1.6% over conventional methods. Our code is publicly available at https://github.com/497662892/intro_to_ds_project.

1 Introduction

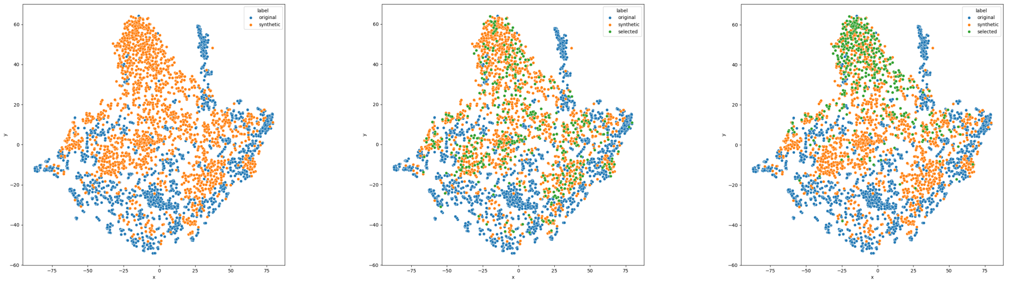
Polyp segmentation models are crucial for assisting doctors during endoscopic examinations, a key procedure in colon cancer screening [5]. Yet, models trained on a single dataset often struggle to generalize well to new domains, due to differences in endoscopic equipment and patient demographics [5, 1, 15, 7]. To address this, researchers have explored various domain generalization approaches, including data augmentation-based techniques, which have proven both simple and effective [16, 4, 8].

While recent efforts have focused on improving the quality of synthetic images [3, 2, 12], less attention has been given to the systematic selection of these samples. Existing methods usually adopt one of two straightforward strategies: random selection [12, 3, 2, 17], which can overlook out-of-distribution (OOD) regions (Figure 1b), or prediction-based selection [13, 8], which can oversample OOD examples and cause distribution shifts (Figure 1c). Both approaches fail to fully leverage synthetic data's potential for expanding feature diversity and enhancing model robustness.

We propose a sampling strategy that jointly considers the distributions of both original and synthetic data. At each iteration, our approach selects the synthetic sample that maximizes the sum of its average distance in feature space to all real samples and previously chosen synthetic samples. This ensures a more balanced coverage of the feature space, avoiding excessive focus on either in-distribution or OOD samples.

Our experimental results demonstrate that this distribution-aware sampling method significantly improves the generalization performance of polyp segmentation models. It outperforms both random and prediction-based selection strategies, without adding substantial computational overhead. ‘

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.

2 Method

2.1 Preliminary

2.1.1 Polyp segmentation

Polyp segmentation is a key task in medical image analysis, aiming to identify polyps in endoscopic images to reduce missed diagnoses [5] (Figure 2). In this study, we measure model performance using the Dice Coefficient [1], a metric that quantifies the overlap between predicted and ground-truth segmentation masks:

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

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.

2.1.2 Random sampling strategy

Algorithm 1 outlines the random sampling strategy. It selects synthetic samples (S_{syn}) to form S_{select}, which are then combined with the original dataset (S_{ori}) to create a new training set (S_{train}) for further training. The distribution of S_{select} is determined solely by the distribution of S_{syn}.

Algorithm 1 Random Sampling Strategy for Data Augmentation

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

2.1.3 Prediction-based sampling strategy

Algorithm 2 describes the prediction-based sampling strategy, which uses a model f_θ trained on the original dataset (S_{ori}). It predicts the performance of samples in the synthetic dataset (S_{syn}) and ranks them by a specified metric. Samples with lower predicted performance are prioritized for inclusion in the augmented training set (S_{train}).

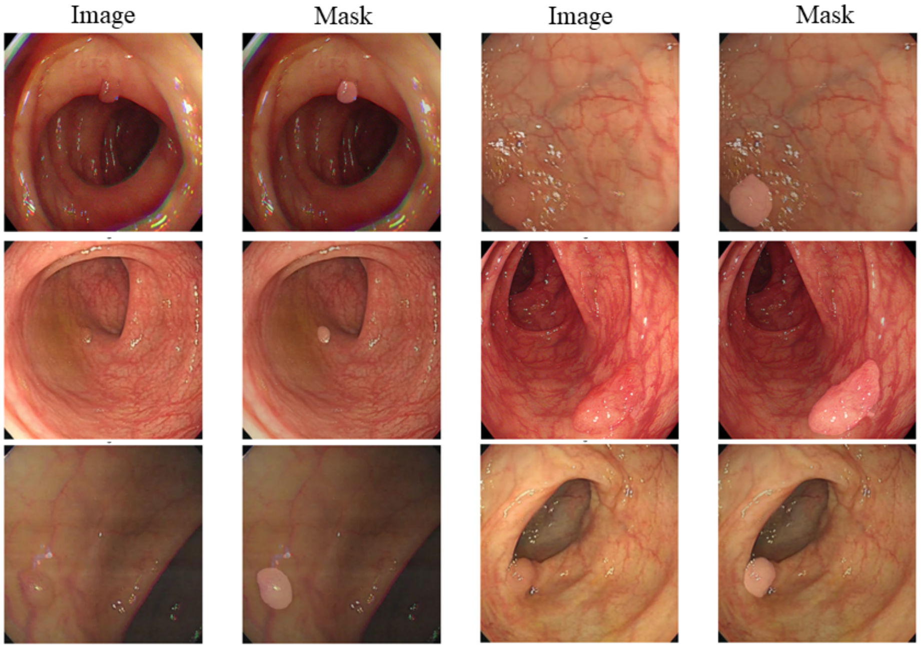


Figure 2: Examples of polyp segmentation.

Algorithm 2 Prediction-Based Sampling Strategy

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

This strategy aims to enhance the training dataset by including samples that the model f_θ finds challenging, focusing primarily on out-of-distribution (OOD) samples related to the original distribution.

2.2 Distribution-Aware Sampling Strategy

We propose a synthetic image selection method that considers the distributions of both real and selected synthetic samples (Algorithm 3). First, we identify the synthetic image with the greatest mean distance to the original training data S_{ori}. Then, for each subsequent selection, we choose the synthetic image that maximizes the sum of its mean distance to S_{ori} and its average distance to the previously selected synthetic samples S_{selected}. This process continues until reaching the desired number of synthetic images.

To simplify computations for high-dimensional data, we extract 512-dimensional features from the last layer of the encoder of a polyp segmentation model trained on S_{ori}, using global average pooling. We use L2 distance in this 512-dimensional space to measure sample similarity.

3 Experiment Setup

3.1 Polyp Segmentation Datasets

We 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

Algorithm 3 Our Synthetic Data Sampling Algorithm

```
Input:
$S_{\text{ori}}$: the set of original data
$S_{\text{syn}}$: the set of synthetic data
$max_{\text{sample}}$: 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:
$S_{\text{select}}.\text{add}(\arg\max_{s \in S_{\text{syn}}} \{D_{\text{syn\_to\_ori}}[s].\text{mean}()\})$    ▷ maximize distance to $S_{\text{ori}}$
Loop:
while $\text{len}(S_{\text{select}}) < max_{\text{sample}}$ 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 $S_{\text{remain}}$
    $D_{\text{re\_to\_sele}}$ = $D_{\text{syn\_to\_syn}}[S_{\text{remain}}, S_{\text{select}}]$    ▷ indexing with $S_{\text{remain}}, S_{\text{select}}$
    $S_{\text{selected}}.\text{add}(\arg\max_{s \in S_{\text{remain}}} \{D_{\text{re\_to\_ori}}[s].\text{mean}() + D_{\text{re\_to\_sele}}[s].\text{mean}()\})$
end while
return $S_{\text{selected}}$
```

[14] (300 images). To expedite the validation of our method, we sampled every 10th frame from each video in the SUN-SEG dataset while preserving the original partitioning, resulting in a training set (1,993 images), validation set (1,442 images), and test set (1,574 images).

The SUN-SEG training set (S_{ori}) was used for model training, the SUN-SEG validation set for model selection, and the remaining datasets served as the testing set to evaluate the generalization performance of the trained model.

3.2 Data Augmentation Method

We used PolypInpainter, a Stable Diffusion-based inpainting model, for data augmentation in this project. PolypInpainter effectively generates realistic inpainted polyps on diverse negative backgrounds while producing high-quality pseudo-masks [9]. Examples are shown in Figure 3.

Negative images from SUN-SEG (781 images) and LD-PolypVideo [10] (1,220 images) were used as backgrounds for inpainting. Following the method and generation settings from PolypInpainter[9], we created 1,993 synthetic images, of which 1,568 were qualified, forming the synthetic dataset S_{syn}.

For this project, we set the target number of selected samples at 400. We selected S_{select} using three strategies: random sampling, prediction-based sampling, and our proposed sampling method. These selected samples were combined with the original training set (S_{ori}, i.e., the SUN-SEG training set) respectively to create three different training datasets (S_{train}).

3.3 Segmentation model

We adapted the polyp-PVT model, a robust method for polyp segmentation [1], as our segmentation framework. The baseline model was trained on the SUN-SEG training set (S_{ori}), while the augmented model was trained on the combined training set (S_{train}). All models were trained following the protocols outlined by Dong et al. [1], with the best-performing model during validation proceeding to the testing phase.

To evaluate generalization performance, we tested the models on four datasets: one internal (SUN-SEG test set) and three external (Kvasir-SEG, CVC-ClinicDB, and CVC-ColonDB), with Dice coefficient as performance metric.

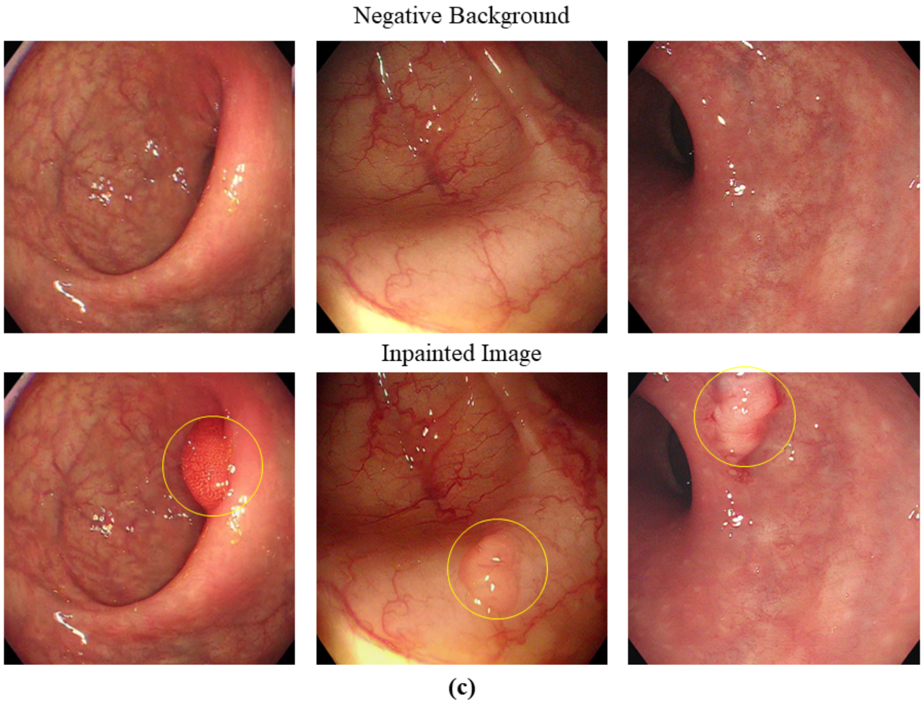


Figure 3: Examples of synthetic endoscopic images.

4 Result

4.1 Generalization in polyp segmentation

Table 1 summarizes the performance of the polyp-PVT model across various datasets with and without data augmentation, using different sampling strategies. Data augmentation with PolypInpainter significantly improved performance on both the SUN-SEG dataset and external datasets. As shown in the last column of Table 1, models trained with data augmentation achieved a consistent increase in overall mDice: +1.3% with random sampling (0.8448 vs. 0.8313), +0.3% with prediction-based sampling (0.8343 vs. 0.8313), and +1.7% with our proposed sampling method (0.8507 vs. 0.8313).

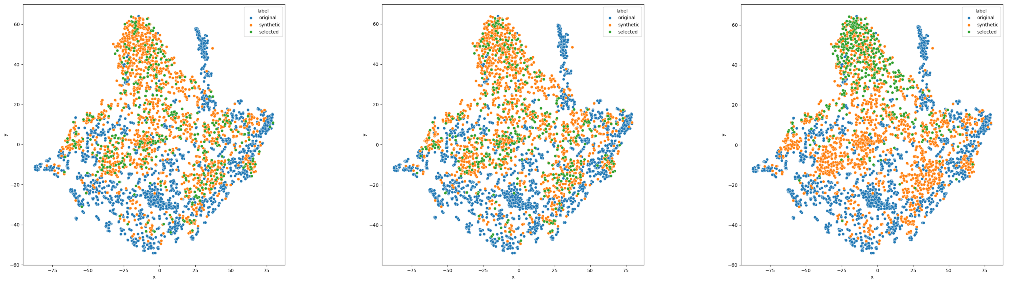
Notably, our proposed sampling strategy not only surpassed random sampling (+0.6%, 0.8507 vs. 0.8448) and prediction-based sampling (+1.6%, 0.8507 vs. 0.8343) in overall mDice but also achieved the best performance across all external datasets (columns 3–5). These results highlight the effectiveness of our sampling method in enhancing model generalization.

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

In Figure 4, we compare the sample distributions produced by different synthetic selection strategies. Compared to random selection (Figure 4a), our method captures more out-of-distribution (OOD) samples (Figure 4b), thereby enhancing feature space diversity. Moreover, compared to the prediction-based selection method (Figure 4c), our strategy avoids oversampling in OOD and maintains the overall integrity of the feature space distribution.



(a) Random selection

(b) Ours

(c) Prediction based selection

Figure 4: The t-SNE visualization of different synthetic data sampling methods.

4.3 Sampling efficiency of our method

The computational complexity per step of our synthetic sample selection algorithm is $O(N \times T)$, where N is the total number of synthetic samples and $T = |S_{\text{select}}|$ is the number of currently selected samples. The overall complexity is bounded by $O(N^3)$. In our experiments, sorting 1,993 samples took approximately 10 seconds on a single CPU in a Jupyter Notebook environment.

5 Conclusion

In this project, we introduced a synthetic sample selection strategy that can aware both the distribution of real data and previously selected synthetic samples. We demonstrated its computational feasibility and superiority over traditional random selection and prediction-based methods for enhancing model generalization in polyp segmentation.

However, our work has several limitations. First, experiments were limited to a single data augmentation method, a fixed target sample size, one polyp segmentation model, one distance metric, and four datasets, suggesting the need for more extensive evaluations. Second, our sampling process still incurs $O(N^3)$ computational complexity, which may need further optimization for large-scale applications. Finally, the applicability of our approach beyond polyp segmentation remains an open question.

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] Jiajian Ma, Fangqi Lu, Silin Huang, Song Wu, and Zhen Li. Generalize polyp segmentation via inpainting across diverse backgrounds and pseudo-mask refinement. *arXiv preprint arXiv:2405.12784*, 2024.

[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.