Spatially Exclusive Pasting: A General Data Augmentation for the Polyp Segmentation

Lei Zhou
zlzhou77@gmail.com

Abstract—Automated polyp segmentation technology plays an important role in diagnosing intestinal diseases, such as tumors and precancerous lesions. Previous works have typically trained convolution-based U-Net or Transformer-based neural network architectures with labeled data. However, the available public polyp segmentation datasets are too small to train the network sufficiently, suppressing each network’s potential performance. To alleviate this issue, we propose a universal data augmentation technology, called SEP, to synthesize more data from the existing datasets. Specifically, we paste the polyp area with surroundings into the same image’s background in a spatial-exclusive manner to obtain a combinatorial number of new images. SEP is a pre-processing method that can be incorporated into any existing network. For a fair comparison, we design a modular platform to facilitate conducting controlled experiments on various networks and datasets. Extensive results show that the proposed method enhances the data efficiency and achieves consistent improvements over baselines. Finally, we hit a new state of the art in this task.

Index Terms—Polyp Segmentation; Data Augmentation; Convolution; Transformer;

I. INTRODUCTION

Colorectal cancer is one of the lethal cancers [1] in the world. Standard colonoscopy relies on the subjective judgment of the physician due to the complex human body condition, like the complicated topology of organs, deformation of organs, and stool. Semantic segmentation [2]–[4] is a fundamental computer vision task by which organs can be clearly segmented at the pixel level. Polyp segmentation, an important branch of semantic segmentation, holds an essential place in the process of colonoscopy. Automated polyp segmentation with deep learning helps doctors to reduce missed cases and greatly decreases the heavy workload of doctors.

Motivated by the huge demand in the real world, a lot of work [5]–[7] has been done on how to better segment polyps. With the aid of pixel-wise labeled data, previous works devoted to designing more and more complex network architectures, like the convolution-based U-Net [8]–[12] and Transformer-based network [13]–[16], to learn better representations for the input images.

These networks are often data-hungry. However, there only exist small datasets for this task since labeling polyp area relies heavily on experienced experts. Thus, proper data augmentation technology impacts the results significantly. Although many methods, like Random Flip, Random Rotate, and Random Zoom, are useful, they are designed for general purposes and not specific for polyp segmentation. An algorithm that could actively locate the polyp area of the picture for data augmentation might be more effective. Copy-Paste [17] improves the instance segmentation task [18] by copying the instance in one image and randomly pasting it to the other image to obtain augmented data. Although this method seemingly meets most of our requirements, there still exist two critical differences between polyp segmentation and instance segmentation.

First, the degree of correlation between the foreground and background is different. For instance segmentation, the objects are more self-identifiable and less relevant to the background. However, the foreground and background of pictures in polyp segmentation have a similar appearance. The determination of polyp needs to be judged by the location and shape of the organ. That is to say, the foreground of polyp must be observed by along with the corresponding background context. Thus, the two core components in Copy-Paste, e.g., only copying foreground and pasting across different images, are not applicable in polyp segmentation. Because they will contribute negatively to build effective contextual relations in

Fig. 1. An example of the augmentation results. The shadows in the augmented GT indicate the locations to paste the cropped polyp.
the polyp segmentation task.

Second, the number of foreground categories is different between instance segmentation and polyp segmentation. A typical instance segmentation dataset, COCO [19], has 80 classes foreground, while the polyp segmentation only has 1 class. For instance segmentation, the various objects being pasted overlappedly can generate convex combined samples to enhance the linearity in feature space and thus make the representation smoother. However, for polyp segmentation, the only foreground category is polyp mass, which makes it impossible to generate a convex combination of different categories in this task. Furthermore, the overlap in the foreground would damage the original appearance of the polyp area, which confuses the learning process. This motivates us to avoid overlapped pasting.

To address the above problems, we propose a novel data argumentation method for polyp segmentation, called Spatially Exclusive Pasting (SEP). The key to our approach lies in self-pasting. Specifically, we first copy the polyp area along with its neighboring background and then find a proper location to paste the duplicate in the same image. The location is determined in a spatially exclusive manner, which avoids corrupting the original foreground area. The proposed method can produce a combinatorial number of new images as well as handle the above two issues. First, the preservation of the surroundings helps to model the relationship between objects and the original background. Second, only pasting in the same images helps to keep a consistent global context. Third, the exclusive manner means that no overlap in the foreground will be introduced after the pasting, reducing the risk of damaging the original polyp appearance.

To verify the effectiveness and generalization of SEP, we construct a modular software platform to facilitate conducting controlled experiments. Extensive experiments on the public benchmark and the representative segmentation architectures show that SEP can increase the data efficiency significantly and can adapt to different types of networks. The main contributions of our work are as follows.

- We propose a universal data augmentation method for better exploiting the existing polyp segmentation datasets.
- We provide a software platform to compare different networks in various datasets under fair settings.
- Extensive experiments are conducted to verify the effectiveness of our method.

II. RELATED WORK

A. Polyp Segmentation

Polyp segmentation has been studied by several pioneering works. In the early stage, UNet [20] was proposed as a universal segmentation architecture to produce full-resolution pixel-wise semantic predictions. Although very simple, it achieved remarkable results on polyp segmentation and formed the basis of nearly all of the following segmentation networks. To enhance the model capacity, ResUNet [8] proposed to extract features using the ResNet [21]. However, the limited dataset hinders the potential of the powerful backbone. Thus, some work proposed to inject prior knowledge into the network. Considering the boundary area is error-prone in polyp segmentation, PraNet [9] designs parallel reverse attention to focus on the edge. ACSNet [10] adaptively selects context to enhance the polyp representation. To better leverage the dataset, DCRNet [12] firstly introduced the interaction across different images through a memory bank. The revolution from Transformer [22], [23] brings new blood for this task. SSFormer [13] proposes to extract the feature using a Transformer block to build long-range dependency and decode the feature in a stepwise manner. To make up for the local inductive bias, FCBFormer [14] designs a two-branch architecture, one using convolution and the other using Transformer. The combination is successful and pushes the art of polyp segmentation to a new high. Conclusively speaking, the backbone for polyp segmentation becomes more and more complex. However, the shortage of datasets makes the performance difference among these networks obscure.

B. Data Augmentation

In contrast to numerous previous works focusing on studying complex network architectures and designing diverse segmentation structures, the research on data argumentation gets relatively less attention. According to the augmentation style, these methods can be grouped into three categories.

1) Geometric Transformation: Geometric transformations are easy to implement and often applied as basic components. LeCun et al. [24] explore assorted affine transforms, like scaling, and horizontal as well as vertical translation, to train LeNet5. Wu et al. [25] utilize various image augmentation approaches, including flipping, lens distortion, cropping, and rotation, on the ImageNet dataset. Although these geometric transformations have decent results on natural image datasets, such as CIFAR-10 and ImageNet, they are more general-purpose and not designed specifically for polyp segmentation.

2) Learning-based Data Augmentation: This branch adopts extra learning parameters to generate new samples satisfying the original data distribution. Smart Augmentation [26] creates a network to generate augmented data to increase the accuracy of the network by merging two or more samples into one class. A target network is trained by the merged images. Meanwhile, the loss of the target network can guide the process of argumentation. However, the algorithm requires a large amount of extra computation and data, which is impractical for polyp segmentation. Another related approach uses adversarial samples to expand training data. Generative Adversarial Networks (GANs) [27] has been used to enhance the generalization of augmented data [28]–[30]. Although these methods are useful in a large dataset, the introduction of a great deal of training time and convergence difficulty indicates that GANs are not suitable for polyp segmentation.

3) Mixing- or Drop-based Data Augmentation: Some works synthesize training data in a non-parametric manner. The core idea is to mix samples to generate a convex combination or randomly drop patches to enhance the robustness of the feature. Mixup [31] is a widely used augmentation that
generates a new image by linearly combining two input images and supervises the new image with the linearly combined labels. CutOut [32] introduces randomly removing square areas of input images by partial occlusion of samples during training. CutMix [33] cuts patches and pastes them on other training data. Although Mixup and its variations have achieved promising results, they are designed for classification in nature, and their procedures are object-agnostic, which violates the object awareness in the segmentation task. Copy-Paste [17] improves the instance segmentation task [18] by extracting object instances and then pasting them on another image to obtain augmented data. This technique is object-aware since it copies the pixels of a concrete object instead of all the pixels in a detection bounding box. This is the most similar method to SEP, but there are three critical differences: 1. we copy and paste the foreground along with the surrounding; 2. the copying and pasting are carried out in the same image; 3. rather than randomly choosing pasting location, we determine the target area using a potential map guided strategy. We will show in the experiments section that the three designs contribute to a better performance significantly.

III. METHODOLOGY

In this section, we introduce the proposed Spatially Exclusive Pasting (SEP) and how to construct a software platform to test SEP’s performance.

A. Spatially Exclusive Pasting

For a given image \( I \in \mathbb{R}^{H \times W \times 3} \) and the corresponding label map \( L \in \{0, 1\}^{H \times W} \), we denote the cropped foreground patch as \( I_f \) and its label map as \( L_f \). Suppose that

\[
\begin{align*}
    h_s, w_s &= \arg \min(L == 1), \\
    h_e, w_e &= \arg \max(L == 1),
\end{align*}
\]

where 1 in the label map indicates the polyp, we have

\[
I_f = I[h_s:h_e, w_s:w_e], \quad (3)
\]

\[
L_f = L[h_s:h_e, w_s:w_e]. \quad (4)
\]

The basic idea of the proposed method for generating new images is very simple: copy the polyp area with surroundings and paste it into other areas in the same image to obtain a combinatorial number of new images. The key concern is where to paste. A trivial solution is to pick locations randomly. However, it may cause conflict in the original foreground area. A cumbersome solution is to calculate the farthest location to the polyp area to avoid overlap. But such manner would lack randomness and introduces incorrect polyp distribution bias. Thus, we design a potential map-based process to raise candidate locations. The whole augmentation process consists of three modules: a potential map generation module to initialize each coordinate with a potential, a pasting module and an update module to update the potential of each coordinate. The whole process is depicted in Fig 2.

First, the potential map generation module randomly initializes a map \( M \in \mathbb{R}^{H \times W} \) of the same size as the input image. The coordinate \((p_x, p_y)\) with the lowest value will be raised to be the candidate location for pasting. We manually set the locations corresponding to the foreground to a large value \( T \), and thus the foreground area is kept from being raised.

Then, we tentatively paste the cropped foreground to the candidate area \( c = [p_x : p_x + h_e - h_s, p_y : p_y + w_e - w_s] \). If there is an overlap of polyp between the cropped patch and the pasting location, we revoke the pasting and lift the potential value of this location to the large value \( T \). If no overlaps, we replace the candidate area with the following convex combination:
Algorithm 1 The process of SEP.

Input:
- Image: \( I \in \mathbb{R}^{H \times W \times 3} \)
- Label map: \( L \in \{0, 1\}^{H \times W} \)
- Potential map: \( M \in \mathbb{R}^{H \times W} \), (random initialized);
- A maximum number: \( T = 10000 \)

Output:
- Augmented Image \( I' \in \mathbb{R}^{H \times W \times 3} \)

1: Crop the poly patch \( I_f \) and \( L_f \) with surroundings using Equation 3;
2: repeat
3: Select the position \((x, y)\) with minimum potential value in \( M \);
4: Tentatively paste and detect overlap;
5: if overlap exists then
6: Revoke the pasting and set \( M(x, y) = T \);
7: else
8: Paste \( I_f \) to \((x, y)\) according to Equation 5 to obtain \( I' \);
9: Update the Potential map according to Equation 7
10: end if
11: Update the Potential map \( M \) with Equation 8;
12: until 10 times
13: return \( I' \);

\[
I[c] = \alpha \times I[c] + (1 - \alpha) \times I_f, \quad (5) \\
L[c] = \alpha \times L[c] + (1 - \alpha) \times L_f, \quad (6)
\]

where \( \alpha = 0.7 \). After the pasting, the potential map is updated by

\[
M[c] = T. \quad (7)
\]

After the tentative pasting, some potential values are changed to a large value \( T \), which means they will never be selected as candidates. Based on a reasonable assumption that the nearby coordinates share a similar likelihood of avoiding overlap, we spread the potential value of each coordinate to its neighbors with the following convolution operation:

\[
M = \text{conv}(M, w), \quad (8)
\]

where \( w \) is a mean filter, e.g.,

\[
w = \begin{bmatrix}
1/9 & 1/9 & 1/9 \\
1/9 & 1/9 & 1/9 \\
1/9 & 1/9 & 1/9
\end{bmatrix} \quad (9)
\]

The pasting module and the update module of the potential map are implemented iteratively 10 times. Thus the foreground area may appear in multiple locations, and an example is depicted in Fig. 1. The full process are summarized in Algorithm 1.

B. The Framework

Since previous works adopt different training policies and tricks to train the network, it’s hard to compare their performance under fair settings. To address this issue, we provide a modular implementation of polyp segmentation. This implementation supports many existing networks and datasets. It’s very convenient to conduct controlled experiments. A basic automatic polyp segmentation project consists of three components: the data loader, the model and the optimizer. The full framework is depicted in Fig. 3. In the data loader, we adopt a pre-specified random seed to control the training/validation/testing phase split, the sequence of loading data and the random data augmentation. As to the optimizer, we carefully adjust the initial learning rate and its scheduler, the max epoch and the gradient clipping settings to ensure all models can hit their reported results. Each model is trained independently with 3 random seeds to reach statistical reliability. For simplicity, we choose \( \{0, 1, 2\} \) as networks’ random seeds.

IV. Experiments

To verify the effectiveness of the proposed methods, we conduct experiments on the different baseline models in this section. To further explore its characteristics, we also provide ablation studies.

A. Implementation Details

All of the experiments are executed on a server with one 2080Ti GPU. The whole project is implemented using PyTorch. We train every model for 150 epochs with an initial learning rate of 1e-4. During training, the learning rate is scheduled by the CosineAnnealingLR [34], with the first 5 epochs to warm up. We adopt gradient clipping to restrict the norm of the gradient to be lower than 20, which stabilizes the training process greatly. And the network parameters are updated with the default Adam optimizer [35]. We still utilize some geometrical data augmentation in our experiments,
including the random horizontal and vertical flip, random rotation and random zoom. We will show that our proposed augmentation is additive to these basic augmentations.

B. Datasets

Following the common practice, we conduct experiments on three public polyp datasets, the Kvasir-SEG [36], EndoScene [37] and CVC-ClinicDB [38]. They have 1000, 912 and 612 labeled images, respectively. EndoScene [37] has been partitioned into train/validation/test subset officially. As to the Kvasir-SEG, prior works [13], [14] randomly divide it into train/validation/test subset using the rate of 8/1/1. However, according to our experiments, the minor differences in data partition cause great gaps in the evaluation results. Therefore we specify three random seeds \{0, 1, 2\} to split the dataset with sklearn. To further ensure reproducibility in different platforms, we also adopt a fixed style that sorts the samples according to file names and extracts the first 80% samples as the training sets, the last 10% as the validation set, and the remainings as the testing set. So, for each model in each dataset, we will provide 4 results, e.g. results for partition \{0, 1, 2, fix\}. For CVC-ClinicDB, we adopt the fixed partition.

C. Evaluations

We report two convincing metrics: the mDice and mIoU. The results of every baseline are obtained through reproductions on our machine. Thanks to the stable experimental settings, all the baselines achieve better results than their reported ones with a certain random seed. With the aid of the proposed augmentation, they all achieve consistent improvements over baselines under all the random seed settings.

1) Baselines: Our method is tested on various representative baselines. They cover the key evolutionary stages in this task.
- ResUNet [8]: Convolution-based;
- PraNet [9]: Convolution + Attention Mechanism;
- SSFormer-S [13]: Transformer-based;
- FCBFormer [14]: Convolution + Transformer.

2) Results on Kvasir-SEG: Table I shows the results of Kvasir-SEG in distinct backbones using assorted random seeds. Since the variances among different seeds are too big to ignore, we choose not to merge them into a “mean-variance” style. Our method boosts all the baselines under every random seed. This verifies that our method effectively improves the data efficiency and well adapts to different types of backbones. Meanwhile, we find that, with proper training, the gap between the newest and the traditional models is rather limited. Qualitative results are illustrated in Fig 4.

3) Results on EndoScene and CVC-ClinicDB: To further demonstrate the generality of SEP, we conduct experiments on EndoScene and CVC-ClinicDB. Since EndoScene [37] provides the official partition, we directly adopt it. For CVC-ClinicDB, we adopt the fixed style to partition the train/val/test split. And we specify a fixed random seed, e.g. 0, during the data augmentation. Table II and Table III show that our methods balance the performance of the different networks.

### Table I

| Seed | Models | mDice(%) ↑ | mIoU(%) ↑ | Seed | Models | mDice(%) ↑ | mIoU(%) ↑ |
|------|--------|------------|------------|------|--------|------------|------------|
| 0    | ResUNet [8] w/ SEP | 92.91 (+0.59)93.5 (+0.71)88.92 | 92.04 (+0.63)92.67 (+0.88)88.21 |
|      | PraNet [9] w/ SEP | 93.02 (+0.74)93.76 (+0.71)89.39 | 92.74 (+0.26)92.63 (+0.41)88.49 |
|      | SSFormer-S [13] w/ SEP | 94.18 (+0.31)94.49 (+0.42)90.55 | 93.72 (+0.75)94.57 (+0.68)90.85 |
|      | FCBFormer [14] w/ SEP | 94.32 (+0.86)94.97 (+1.16)91.07 | 93.13 (+0.23)93.13 (+0.36)89.46 |
| 1    | ResUNet [8] w/ SEP | 90.49 (+0.05)90.54 (+0.14)85.23 | 92.72 (+0.4)92.67 (+0.48)88.89 |
|      | PraNet [9] w/ SEP | 91.35 (+0.23)91.68 (+0.36)86.7 | 93.1 (+0.82)92.63 (+0.73)88.46 |
|      | SSFormer-S [13] w/ SEP | 92.31 (+0.45)92.76 (+0.68)86.51 | 93.19 (+0.65)93.84 (+0.72)89.77 |
|      | FCBFormer [14] w/ SEP | 91.93 (+0.53)92.46 (+0.79)87.52 | 93.91 (+0.2)94.11 (+0.18)90.02 |

![Fig. 4. Error map visualization. The white area is obtained by judging whether the prediction equals the ground truth. Our method improves the boundary areas.](image-url)
TABLE II
Results on ENDOScene.

| Methods            | mDice(%) ↑ | mIoU(%) ↑ |
|--------------------|------------|-----------|
| ResUNet [8]        | 86.02      | 79.34     |
| w/SEP              | (+0.2)87.22| (+1.0)80.64|
| PraNet [9]         | 86.3       | 79.6      |
| w/SEP              | (+0.1)86.4 | (0.5)80.1 |
| SSFormer-S [13]    | 86.95      | 80.6      |
| w/SEP              | (+0.3)87.2 | (+1.0)80.85|
| FCBFormer [14]     | 88.1       | 82.1      |
| w/SEP              | (+0.2)88.3 | (+0.2)82.3|

TABLE III
Results on CVC-ClinicDB.

| Methods            | mDice(%) ↑ | mIoU(%) ↑ |
|--------------------|------------|-----------|
| ResUNet [8]        | 79.3       | 70.7      |
| w/SEP              | (+3.7)83.0 | (+3.8)74.5|
| PraNet [9]         | 76.9       | 68.4      |
| w/SEP              | (+3.2)80.1 | (+3.0)71.4|
| SSFormer-S [13]    | 83.5       | 74.4      |
| w/SEP              | (+0.7)84.2 | (+1.7)76.1|
| FCBFormer [14]     | 81.6       | 72.0      |
| w/SEP              | (+1.5)83.1 | (+1.4)73.4|

D. Ablation Study

To further explore the characteristics of the method, we ablate the pasting process carefully. First, we study the pasting location. We replace the potential map-based pasting with random pasting or random non-overlap pasting. Random pasting means randomly picking the pasting location, and the random non-overlap pasting adds a restriction to forbid pasting in the foreground area. We try to paste the foreground only as non-overlap pasting adds a restriction to forbid pasting in the foreground area. Then we investigate the effect of cross-frame pasting. We try to paste the foreground only as non-overlap pasting adds a restriction to forbid pasting in the foreground area. We attempt to conduct cross-frame pasting. Then we investigate the effect of these pasting styles. Results in Table IV show that our potential map-guided strategy outperforms all the other variants. We attribute this to reducing spatial conflict among the polyp areas and keeping the original context.

TABLE IV
Ablations on the pasting manners.

| Configs              | mDice(%) ↑ | mIoU(%) ↑ |
|----------------------|------------|-----------|
| No Paste             | 93.19      | 89.05     |
| Random Paste         | 84.31      | 78.15     |
| Non-overlap Paste    | 93.28      | 88.86     |
| Foreground-Only Paste| 93.39      | 89.12     |
| Cross-Frame Paste    | 93.37      | 89.03     |
| Spatially Exclusive Paste| 93.84 | 89.77 |

V. Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.

VI. Conclusion

In this work, we propose a novel data augmentation strategy, called SEP, to alleviate the shortage of data for the polyp segmentation task. SEP searches the spatially exclusive location to conduct intra-frame pasting. We conduct extensive controlled experiments under a fair platform to verify that the proposed method consistently improves the performance of different networks. In the future, we will work on extending this method to more tasks and explore how to learn from various data sources.

REFERENCES

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, Ahmedin Jemal, et al., “Globocat estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” Ca Cancer J Clin, vol. 68, no. 6, pp. 394–424, 2018.
[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3341–3440.
[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.
[4] Hengshuang Zhao, Jianping Shi, Xiaojian Qi, Xiaogang Wang, and Jiaya Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[5] Nima Tajbakhsh, Suryakant R Gurudu, and Jianming Liang, “Automated polyp detection in colonoscopy videos using shape and context information,” IEEE transactions on medical imaging, vol. 35, no. 2, pp. 630–644, 2015.
[6] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Peng Ann Heng, “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos,” IEEE journal of biomedical and health informatics, vol. 21, no. 1, pp. 65–75, 2016.
[7] Ruikai Zhang, Yali Zheng, Carmen CY Poon, Dinggang Shen, and James YW Lau, “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker,” Pattern recognition, vol. 83, pp. 209–219, 2018.
[8] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang, “Road extraction by deep residual u-net,” IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753, 2018.
[9] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jiabing Shen, and Ling Shao, “Planet: Parallel reverse attention network for polyp segmentation,” in International conference on medical image computing and computer-assisted intervention. Springer, 2020, pp. 263–273.
[10] Ruifei Zhang, Guanlin Bi, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu, “Adaptive context selection for polyp segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 253–262.
[11] Abhishek Srivastava, Debesh Jha, Sukalpa Chanda, Umapada Pal, Håvard D Johansen, Dag Johansen, Michael A Riegler, Sharib Ali, and Pål Halvorsen, “Msrf-net: A multi-scale residual fusion network,” in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, pp. 8665–8672.
[12] Jinfeng Wang, Qiming Huang, Feilong Tang, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2021, pp. 3431–3440.
[13] Edward Sanderson and Bogdan J Matuszewski, “Fcn-transformer feature fusion for polyp segmentation,” in Annual Conference on Medical Image Understanding and Analysis. Springer, 2022, pp. 892–907.
[14] Nguyen Thanh Duc, Nguyen Thi Oanh, Nguyen Thi Thuy, Tran Minh Triet, and Viet Sang Dinh, “Colonformer: An efficient transformer based method for colon polyp segmentation,” IEEE Access, vol. 10, pp. 80575–80586, 2022.
[15] Yunfeng Zhang, Huiye Liu, and Qiang Hu, “Transpose: Fusing transformers and cnns for medical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 14–24.
Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

Aashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville, “A benchmark for endoluminal scene segmentation of colonoscopy images,” Journal of healthcare engineering, vol. 2017, 2017.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” Computerized medical imaging and graphics, vol. 43, pp. 99–111, 2015.