Automatic Polyp and Instrument Segmentation in MedAI-2021

Yu-Cheng Chou¹
1. School of Computer Science, Wuhan University

Abstract
Polyp and instrument segmentation plays a vital role in the early diagnosis of colorectal cancer (CRC), which is usually performed by visual inspection of the digestive system with an endoscope to identify polyps. However, recent works only focus on the accuracy of prediction in the positive samples while omitting the False-Positive (FP) predictions in the negative samples that might mislead the physicians. Here, we propose a novel Dual Model Filtering (DMF) strategy, which efficiently removes FP predictions in negative samples with metrics based threshold setting. To better adapt to high-resolution input with various distributions, we embed the PVTv2 [1] backbone to the SINetV2 [2] framework as our model. Intuitively, we apply camouflaged object detection (COD) SINetV2 for better identification ability since the polyp segmentation is one of the downstream tasks of. Experiments on the datasets provided during the MedAI [3] challenge demonstrate our method achieves excellent performance. We also conduct extensive experiments to study the effectiveness of the DMF. We find that our method works well under different data distributions, making it a favorable solution for problems where training dataset shares different negative samples distribution comparing to testing dataset.

Keywords: Polyp and instrument segmentation; Colonoscopy

Backbone
During the colonoscopy, multi-scale polyps and instruments move at various speeds, and thus, the limited and fixed receptive field of CNN is inadequate to address this problem [4]. Here, we utilize PVTv2 [1] as the feature extractor in our framework. Note that compared to ViT [5], PVTv2 retrieves only global receptive fields and gets pyramid features as in Res2Net [6], which is more suitable for dense prediction tasks than conventional CNNs.

Loss Function
We apply the deep supervision for the three side-outputs and the global map with a pixel position-aware (PPA) loss function [7]. Compared to the conventional IoU loss and binary cross entropy (BCE) loss that treat all pixels equally and ignore the difference between pixels, the PPA loss function assigns more weights to pixels which are more difficult for identification to emphasize their importance.

Implementation Details
Our training strategy contains three steps:

Data-Oriented Pre-Training
To better fit the backbone with the class-agnostic polyp and instrument segmentation task, we pre-train for different tasks with Kvasir-Polyp [8] (1,000 samples) and Kvasir-Instrument [9] (590 samples) datasets, respectively. We randomly select 100 samples from the original dataset as the validation set and the rest for training during the training stage. The batch size is set to 32, and the learning rate starts at $10^{-4}$, dividing by 10 every 20 epochs into 100 epochs. We conduct the experiments on single NVIDIA RTX TITAN GPU with PyTorch toolbox.

Negative Fine-Tuning
We achieve satisfying results after the data-oriented pre-training in the last stage. However, the model learns with only positive samples and might cause FP prediction for negative ones. Consequently, we fine-tune it for different tasks with additional negative samples that account for one-third of the total data. For polyp segmentation,
we manually select 300 negative samples from Kvasir-Instrument. For instrument segmentation, we select 200 negative samples from Kvasir-Polyp. This stage shares the same settings as in the pre-training.

**Dual Model Filtering**

Applying the model after fine-tuning can remove some FP predictions, but it is still prone to generate ambiguous predictions without a clear edge. To alleviate the FP predictions, we compare the predictions after CRF [10] of the model trained without and with the negative samples. We observe that predictions of the two models vary a lot in negative samples while being consistent in positive ones. As shown in Fig.1, we demonstrate the predictions of the negative samples to reveal the nature of negative fine-tuned model which inevitably generate FP predictions. To eliminate the FP predictions as far as possible, we proposed a new post processing strategy named dual model filtering (DMF) by computing the similarity of the predictions of two model. Specifically, we set the IoU as the metric to measure the similarity. Note that, the original ground truth and prediction of the metric are prediction of model after pre-training and model after fine-tuning, respectively. Here, we set the final predictions are their intersection if above the threshold and O if below the threshold.

**Results and Analysis**

**Metrics**

During evaluation, we use the following metrics: Sørensen-Dice coefficient (Dc), accuracy (Acc), recall (Rcl), precision (Prec), and IoU (also called Jaccard), $S_\alpha$ [11], $E_{\phi}^w$ [12], $F_\beta^w$ [13], and MAE ($\mathcal{M}$).

**Submission**

As shown in Tab.1, we have three submissions in this competition, where the scores are provided by the official evaluation.

<table>
<thead>
<tr>
<th>Table 1: Submission results.</th>
<th>Acc↑</th>
<th>IoU↑</th>
<th>Dc↑</th>
<th>Rcl↑</th>
<th>Prec↑</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polyp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp1</td>
<td>0.970</td>
<td>0.848</td>
<td>0.899</td>
<td>0.919</td>
<td>0.910</td>
</tr>
<tr>
<td>Exp2</td>
<td>0.975</td>
<td>0.833</td>
<td>0.878</td>
<td>0.897</td>
<td>0.885</td>
</tr>
<tr>
<td>Exp3</td>
<td>0.977</td>
<td>0.821</td>
<td>0.861</td>
<td>0.860</td>
<td>0.881</td>
</tr>
<tr>
<td><strong>Instrument</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp1</td>
<td>0.974</td>
<td>0.914</td>
<td>0.951</td>
<td>0.960</td>
<td>0.950</td>
</tr>
<tr>
<td>Exp2</td>
<td>0.990</td>
<td>0.912</td>
<td>0.943</td>
<td>0.969</td>
<td>0.923</td>
</tr>
<tr>
<td>Exp3</td>
<td>0.992</td>
<td>0.912</td>
<td>0.948</td>
<td>0.980</td>
<td>0.923</td>
</tr>
</tbody>
</table>

**Ablation Analysis**

We evaluate the performance of the model trained without negative samples (after pre-training), with negative samples (after fine-tuning), and with DMF. During DMF in the validation dataset, we set the IoU threshold to 0.25 by analyzing the prediction distribution. As shown in Tab.1 and Tab.2, the model trained without negative samples achieves satisfying results. We apply the DMF strategy to eliminate the misaligned predictions to address the negative samples better and eventually achieve improved performance.

<table>
<thead>
<tr>
<th>Table 2: Ablation study for polyp segmentation. We evaluate the proposed methods with the validation dataset that contains negatives samples. ↑ and ↓ denote larger and smaller are better, respectively.</th>
<th>Acc↑</th>
<th>IoU↑</th>
<th>Dc↑</th>
<th>$S_\alpha$↑</th>
<th>$E_{\phi}^w$↑</th>
<th>$F_\beta^w$↑</th>
<th>$\mathcal{M}$↓</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polyp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o Neg</td>
<td>0.953</td>
<td>0.915</td>
<td>0.952</td>
<td>0.933</td>
<td>0.954</td>
<td>0.951</td>
<td>0.041</td>
</tr>
<tr>
<td>w/ Neg</td>
<td>0.961</td>
<td>0.880</td>
<td>0.927</td>
<td>0.923</td>
<td>0.949</td>
<td>0.921</td>
<td>0.034</td>
</tr>
<tr>
<td>w/ DMF</td>
<td>0.980</td>
<td>0.915</td>
<td>0.946</td>
<td>0.949</td>
<td>0.972</td>
<td>0.946</td>
<td>0.015</td>
</tr>
<tr>
<td><strong>Instrument</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o Neg</td>
<td>0.983</td>
<td>0.925</td>
<td>0.957</td>
<td>0.950</td>
<td>0.977</td>
<td>0.949</td>
<td>0.016</td>
</tr>
<tr>
<td>w/ Neg</td>
<td>0.756</td>
<td>0.689</td>
<td>0.752</td>
<td>0.720</td>
<td>0.747</td>
<td>0.751</td>
<td>0.243</td>
</tr>
<tr>
<td>w/ DMF</td>
<td>0.993</td>
<td>0.928</td>
<td>0.953</td>
<td>0.956</td>
<td>0.981</td>
<td>0.949</td>
<td>0.007</td>
</tr>
</tbody>
</table>

**Conclusion and discussion**

In this paper, we proposed a novel DMF strategy to alleviate the FP predictions with metrics based threshold setting under different data distribution. Meanwhile, benefiting from the SINetV2 embedded with PVTv2, the proposed method can adapt to high-resolution camouflage input with various distributions. Ablation studies demonstrate that the model with the DMF strategy outperforms the original one.

This method can be improved by using adaptive threshold or learnable fusion strategy to increase the generality. In addition, it may be beneficial to apply structure measure as the filtering metrics to focus on the structure similarity of two predictions.

**Acknowledgments**

Many thanks to Ge-Peng Ji & Deng-Ping Fan at IIAI CV&Med team for providing technical supports.
References


