A Comprehensive Study on Colorectal Polyp Segmentation With ResUNet++, Conditional Random Field and Test-Time Augmentation

Debesh Jha ⓘ, Pia H. Smedsrud, Dag Johansen ⓘ, Thomas de Lange ⓘ, Håvard D. Johansen ⓘ, Pål Halvorsen, and Michael A. Riegler ⓘ

Abstract—Colonoscopy is considered the gold standard for detection of colorectal cancer and its precursors. Existing examination methods are, however, hampered by high overall miss-rate, and many abnormalities are left undetected. Computer-Aided Diagnosis systems based on advanced machine learning algorithms are touted as a game-changer that can identify regions in the colon overlooked by the physicians during endoscopic examinations, and help detect and characterize lesions. In previous work, we have proposed the ResUNet++ architecture and demonstrated that it produces more efficient results compared with its counterparts U-Net and ResUNet. In this paper, we demonstrate that further improvements to the overall prediction performance of the ResUNet++ architecture can be achieved by using Conditional Random Field (CRF) and Test-Time Augmentation (TTA). We have performed extensive evaluations and validated the improvements using six publicly available datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib Polyp DB, ASU-Mayo Clinic Colonoscopy Video Database, and CVC-VideoClinicDB. Moreover, we compare our proposed architecture and resulting model with other state-of-the-art methods. To explore the generalization capability of ResUNet++ on different publicly available polyp datasets, so that it could be used in a real-world setting, we performed an extensive cross-dataset evaluation. The experimental results show that applying CRF and TTA improves the performance on various polyp segmentation datasets both on the same dataset and cross-dataset. To check the model's performance on difficult to detect polyps, we selected, with the help of an expert gastroenterologist, 196 sessile or flat polyps that are less than ten millimeters in size. This additional data has been made available as a subset of Kvasir-SEG. Our approaches showed good results for flat or sessile and smaller polyps, which are known to be one of the major reasons for high polyp miss-rates. This is one of the significant strengths of our work and indicates that our methods should be investigated further for use in clinical practice.

Index Terms—Colonoscopy, polyp segmentation, ResUNet++, conditional random field, test-time augmentation, generalization.

Debesh Jha is with the SimulaMet, 0167 Oslo and UiT The Arctic University of Norway, 9037 Tromsø (e-mail: debesh@simula.no).

Pia H. Smedsrud is with the SimulaMet, 0167 Oslo and University of Oslo, 0315 Oslo (e-mail: pia@simula.no).

Dag Johansen and Håvard D. Johansen are with the UiT The Arctic University of Norway, 9037 Tromsø (e-mail: dag.johansen@uit.no; havard.johansen@uit.no).

Thomas de Lange is with the Medical Department Sahlgrenska University Hospital – Mölndal, Region Västra Götaland, 431 30 Mölndal, Sweden, Department of Medical Research, Brum Hospital, Vestre Viken, 1346 Gjettum, Norway and also with the Department of Molecular and Clinical Medicine, Sahlgrenska Academy, University of Gothenburg, 40530 Göteborg, Sweden (e-mail: t.de.lange@medisin.uio.no).

Pål Halvorsen is with the SimulaMet and Oslo Metropolitan University, 0167 Oslo (e-mail: paalh@simula.no).

Michael A. Riegler is with the SimulaMet, 0167 Oslo (e-mail: michael@simula.no).

Digital Object Identifier 10.1109/JBHI.2021.3049304

I. INTRODUCTION

CANCER is a primary health problem of contemporary society, with colorectal cancer (CRC) being the third most prevailing type in terms of cancer incidence and second in terms of mortality globally [2]. Colorectal polyps are the precursors for the CRC. Early detection of polyps through high-quality colonoscopy and regular screening are cornerstones for the prevention of colorectal cancer [3], since neoplastic lesions such as adenomas can be found and resected before transforming to cancer and subsequently reducing CRC morbidity and mortality.

Regardless of the achievement of colonoscopy in lowering cancer burden, the estimated adenoma miss-rate is around 6–27% [5]. In a recent pooled analysis of 8 randomized tandem colonoscopy studies, polyps smaller than 10 mm, sessile, and flat polyps [6] are shown to most often be missed [7]. Another reason why polyps are missed may be that the polyp either was not in the visual field or was not recognized despite being in the visual field due to fast withdrawal of the colonoscope [8]. The adenoma miss-rate could be reduced by improving the quality of bowel preparation, applying optimal observation techniques, and ensuring a colonoscopy withdrawal time of at least six minutes [8]. Moreover, adenoma detection rate can also be improved by using advanced techniques or devices, for example, auxiliary imaging devices, colonoscopes with increased field of view, add-on-devices, and colonoscopes with integrated inflatable, reusable balloons [3].
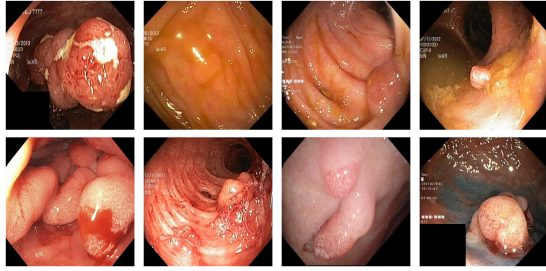
Fig. 1. Example images showing the variations in shape, size, color, and appearance of polyps from the Kvasir-SEG [4].

The structure and characteristics of a colorectal polyp changes over time at different development stages. Polyps have different shapes, sizes, colors, and appearances, which makes them challenging to analyze (see Fig. 1). Moreover, there are challenges such as the presence of image artifacts like blurriness, surgical instruments, intestinal contents, flares, and low-quality images that can cause errors during segmentation.

Polyp segmentation is of crucial relevance in clinical applications to focus on the particular area of the potential lesion, extract detailed information, and possibly remove the polyp if necessary. A Computer-Aided Diagnosis (CADx) system for polyp segmentation can assist in monitoring and increasing the diagnostic ability by increasing the accuracy, precision, and reducing manual intervention. Moreover, it could lead to less segmentation errors than when conducted subjectively. Such systems could reduce doctor's workload and improve clinical workflow. Lumen segmentation helps clinicians navigate through the colon during the examination, and it can be useful to establish a quality metric for the explored colon wall [9]. Thus, an automated CADx system could be used as a supporting tool to reduce the miss-rate of the overlooked polyps.

A CADx system could be used in a clinical setting if it addresses two common challenges: (i) Robustness (i.e., the ability of the model to consistently perform well on both easy and challenging images), and (ii) Generalization (i.e., a model trained on specific intervention in a specific hospital should generalize across different hospitals) [10]. Addressing these challenges is key to design a powerful semantic segmentation system for medical images. Generalization capability checks the usefulness of the model across different available datasets coming from different hospitals and must finally be confirmed in multi-center randomized trials. A good generalizable model could be a significant step toward developing an acceptable clinical system. A cross-dataset evaluation is crucial to check the model on the unseen polyps from other sources and test the generalizability of it.

Toward developing a robust CADx system, we have previously proposed ResUNet++ [1]: an initial encoder-decoder based deep-learning architecture for segmentation of medical images, which we trained, validated, and tested on the publicly available Kvasir-SEG [4] and CVC-ClinicDB [11] datasets. In this paper, we describe how the ResUNet++ architecture can be extended by applying Conditional Random Field (CRF) and Test-Time Augmentation (TTA) to further improve its prediction

performance on segmented polyps. We have tested our approaches on six publicly available datasets, including both image datasets and video datasets. We have intentionally incorporated video datasets from colonoscopies to support the clinical significance. Usually, still-frames have at least one polyp sample. Videos have a situation where frames consist of both polyp and non-polyp. Therefore, we have tested the model on these video datasets and provided a new benchmark for the segmentation task. We have used extensive data augmentation to increase the training sample and used a comprehensive hyperparameter search to find optimal hyperparameters for the dataset. We have provided a more in-depth evaluation by including more evaluation metrics, and added justification for the ResUNet++, CRF, and TTA.

Additionally, we have performed extensive experiments on the cross-data evaluation, in-depth analysis of best performing and worst performing cases, and comparison of the proposed method with other recent works. Moreover, we have pointed out the necessity of solving tasks related to the miss-detection of flat and sessile polyps, and showed that our combining approach could detect the overlooked polyps with high efficiency, which could be of significant importance in the clinical settings. For this, we also released a dataset consisting of sessile or flat polyps publicly. Furthermore, we have emphasized the use of cross-dataset evaluation by training and testing the model with images coming from various sources to achieve the generalizability goal.

In summary, the main contributions are as follows:

1) We have extended the ResUNet++ deep-learning architecture [1] for automatic polyp segmentation with CRF and TTA to achieve better performance. The quantitative and qualitative results show that applying CRF and TTA is effective.

2) We validate the extended architecture on a large range of datasets, i.e., Kvasir-SEG [4], CVC-ClinicDB [11], CVC-ColonDB [12], EITS-Larib [13], ASU-Mayo Clinic Colonoscopy Video Database [14], and CVC-VideoClinicDB [15], [16], and we compare our proposed approaches with the recent state-of-the-art (SOTA) algorithms and set a new a baseline. Moreover, we have compared our work with other recent works, which is often lacking in comparable studies.

3) We selected 196 flat or sessile polyps that are frequently missed during colonoscopy examination [7] from the Kvasir-SEG with the help of an expert gastroenterologist. We have conducted experiments on this separate dataset to investigate how well our model performs on challenging polyps. Moreover, we release these polyp images and segmentation masks as a part of the Kvasir-SEG dataset so that researchers can build novel architectures and improve the results.

4) Our model has better detection of smaller and flat or sessile polyps, which are frequently missed during colonoscopy [7], which is a major strength compared to existing works.

5) In medical clinical practice, generalizable models are essential to target patient population. Our work is focused

on generalizability, previously not much explored in the community. To promote generalizable Deep Learning (DL) models, we have trained our models on Kvasir-SEG and CVC-ClinicDB and tested and compared the results over five publicly available diverse unseen polyp datasets. Moreover, we have mixed two diverse datasets and conducted further experiments on other unseen datasets to show the behaviour of the model on the images captured using different devices.

II. RELATED WORK

Over the past decades, researchers have made several efforts at developing CADx prototypes for automated polyp segmentation. Most of the prior polyp segmentation approaches were based on analyzing either the polyp's edge or its texture. More recent approaches used Convolutional Neural Network (CNN) and pre-trained networks. Bernal *et al.* [11] introduced a novel method for polyp localization that used WM-DOVA energy maps for accurately highlighting the polyps, irrespective of its type and size. Pozdeev *et al.* [17] presented a fully automated polyp segmentation framework using pixel-wise prediction based upon the Fully Convolutional Network (FCN). Bernal *et al.* [18] hosted the automatic polyp detection in colonoscopy videos sub-challenge, and later on, they presented a comparative validation of different methods for automatic polyp detection and concluded that the SOTA CNN based methods provide the most promising results.

Akbari *et al.* [19] used the FCN-8S network and Otsu's thresholding method for automated colon polyp segmentation. Wang *et al.* [20] used the SegNet [21] architecture to detect polyps. They obtained high sensitivity, specificity, and receiver operating characteristic (ROC) curve value. Their algorithm could achieve a speed of 25 frames per second with some latency during real-time video analysis. Guo *et al.* [22] used a Fully Convolutional Neural Network (FCNN) model for the Gastrointestinal Image ANAlysis (GIANA) polyp segmentation challenge. The proposed method won first place in the 2017 GIANA challenge for both standard definition (SD) and high definition image and won second place in the SD image segmentation task in the 2018 GIANA challenge. Yamada *et al.* [23] developed a CADx support system that can be used for the real-time detection of polyps reducing the number of missed abnormalities during colonoscopy.

Poorneshwaran *et al.* [24] used a Generative Adversarial Network (GAN) for polyp image segmentation. Kang *et al.* [25] used Mask R-CNN, which relies on ResNet50 and ResNet101, as a backbone structure for automatic polyp detection and segmentation. Ali *et al.* [26] presented various detection and segmentation methods that could classify, segment, and localize artifacts. Additionally, there are several recent interesting studies on polyp segmentation [27]–[30]. They are useful steps toward building an automated polyp segmentation system. There are also some works which have hypothesized that coupling the existing architecture by applying a careful post-processing technique could improve the model performance [1], [31].

From the presented related work, we observe that automatic CADx systems in the area of polyp segmentation are becoming mature. Researchers are conducting a variety of studies with different designs ranging from a retrospective study, prospective study, to post hoc examination of the prospectively obtained dataset. Some of the models achieve very high performance with smaller training and test datasets [1], [20], [32]. The algorithms used for building the models are the ones that use handcrafted-, CNN- or pre-trained-features from ImageNet [33], where DL based algorithms are outperforming and gradually replacing the traditional handcrafted or machine learning (ML) approaches. Additionally, the performance of the models improves by the use of advance DL algorithms, especially designed for polyp segmentation task or any other similar biomedical image segmentation task. Moreover, there is interest for testing the proposed architectures with more than one dataset [1], [20].

The main drawbacks in the field are the minimal effort applied towards testing the generalizability of the CADx system possible to achieve with the cross-dataset test. Additionally, there is almost no effort involved in designing a universal model that could accurately segment polyps coming from different sources, critical for the development of CADx for automated polyp segmentation. Besides, most of the current works have proposed algorithms that are tested on single, often small, imbalanced, and explicitly handpicked datasets. This renders conclusions regarding the performance of the algorithms almost useless (compared to other areas in ML like, for example, natural image classification or action recognition where the common practice is to test on more than one dataset and make source code and datasets publicly available). Additionally, the used datasets are often not public available (restricted and difficult to access), and the total number of images and videos used in the study are not sufficient to conjecture that the system is robust and generalizable for use in clinical trials. For instance, the model can produce output segmentation maps with high sensitivity and precision on a particular dataset and completely fails on other modality images. Moreover, existing work often use small training and test datasets. These current limitations make it harder to develop a robust and generalizable systems.

Therefore, we aim to develop a CADx based support system that could achieve high performances irrespective of the datasets. To achieve the goal, we have done extensive experiments on various colonoscopy images and video datasets. Additionally, we have mixed the dataset from multiple centers and tested it on other diverse unseen datasets to achieve the goal of building a generalizable and robust CADx system that potentially produces no segmentation errors. Moreover, we set a new benchmark for the publicly available datasets that can be improved in the future.

III. THE RESUNET++ ARCHITECTURE

ResUNet++ is a semantic segmentation deep neural network designed for medical image segmentation. The backbone for ResUNet++ architecture is ResUNet [34]: an encoder-decoder network and based on U-Net [35]. The proposed architecture takes the benefit of residual block, squeeze and excite block [36], atrous spatial pyramid pooling (ASPP) [37], and attention block [38]. What distinguishes ResUNet++ from ResUNet is the use of squeeze-and-excitation blocks (marked in dark gray) at the encoder, the ASPP block, (marked in the dark red) at bridge
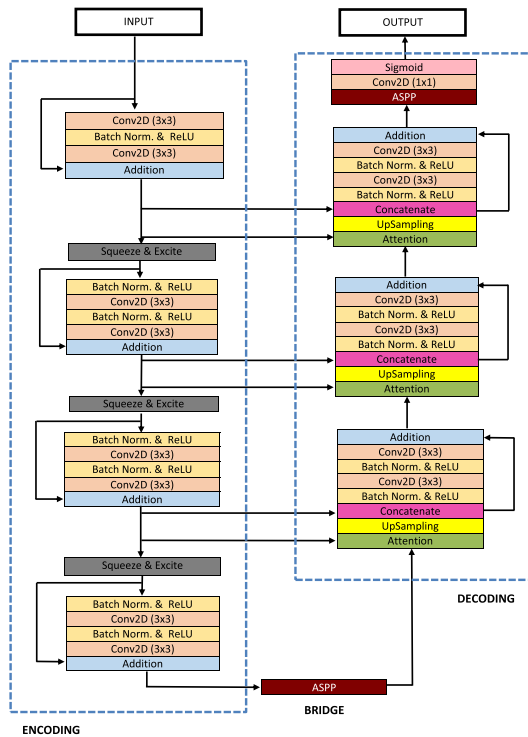
Fig. 2.   ResUNet++ architecture [1].

and decoder, and the attention block (marked in light green) at the decoder (see Fig. 2).

In the ResUNet++ model, we introduce the sequence of squeeze and excitation block to the encoder part of the network. Additionally, we replace the bridge of ResUNet with ASPP. In the decoder stage, we introduce a sequence of attention block, nearest-neighbor up-sampling, and concatenate it with the relevant feature map from the residual block of the encoder through skip connection. This process is followed by the residual unit with identity mapping, as shown in Fig. 2.

We also introduce a series of additional skip connections from the residual unit of the encoder section to the attention block of the decoder section. We assign the number of filters $[32, 64, 128, 256, 512]$, along with the levels in the encoder section, which are the values in our ResUNet++ architecture. These filter combinations achieved the best results in our ResUNet++ experiment. In the decoder section, the number of the filters are reversed, and the sequence becomes $[512, 256, 128, 64, 32]$. As the semantic gap between the feature map of the encoder and decoder blocks are supposed to decrease, the number of filters in the convolution layers of the decoder block is also decreased to achieve better semantic coverage. Through this, we ensure that the overall quality of the feature maps is more alike to the ground truth mask. This is especially important as the loss in semantic space is likely to decrease, and therefore it will become more feasible to find a meaningful representation in semantic space.

The overall ResUNet++ architecture consists of one stem block with three encoder blocks, an ASPP between the encoder and the decoder, and three decoder blocks. All the encoder and decoder blocks use the standard residual learning approach. Skip connections are introduced between encoder and decoder for the

propagation of information. The output of the last decoder block is passed through the ASPP, followed by a 1×1 convolution and a sigmoid activation function. All convolutional layers except for the output layer are batch normalized [39] and are activated by a Rectified Linear Unit (ReLU) activation function [40]. Finally, we get the output as binary segmentation maps. A brief explanation of each block is provided in the following sub-sections.

A. Residual Blocks

Training a deep neural network by expanding network depth can potentially improve overall performance. Nevertheless, simply stacking the CNN layer could also hamper the training process and cause exploding/vanishing gradient when backpropagation occurs [41]. Residual connections facilitate the training process by directly routing the input information to the output and preserve the nobility of the gradient flow. The residual function simplifies the objective of optimization without any additional parameters and boosts the performance, which is the inspiration behind the deeper residual-based network [42]. Equation (1) below shows the working principle.

$$y_n = F(x_n, W_n) + x_n \tag{1}$$

Here, x_n is the input and $F(\cdot)$ is the residual function. The residual units consist of numerous combinations of Batch Normalization (BN), ReLU, and convolution layers. A detailed description of the combinations used and their impact can be found in the work of He *et al.* [43]. We have employed the concept of a pre-activation residual unit in the ResUNet++ architecture from ResUNet.

B. Squeeze and Excitation Block

The squeeze and excitation (SE) block is the building block for the CNN that re-calibrates channel-wise feature response by explicitly modeling interdependencies between the channels [36]. The SE block learns the channel weights through global spatial information that increases the sensitivity of the effective feature maps, whereas it suppresses the irrelevant feature maps [1]. The feature maps produced by the convolution have only access to the local information, meaning they have no access to the global information left by the local receptive field. To address this limitation, we perform a squeeze operation on the feature maps using the global average pooling to generate a global representation. We then use the global representation and perform sigmoid activation that helps us to learn a non-linear interaction between the channels, and capture the channel-wise dependencies. Here, the sigmoid activation output acts as a simple gating mechanism that ensures us to adaptively recalibrate the feature maps produced by the convolution. The adaptive recalibration or excitation operation explicitly models the interdependencies between the feature channels. The SE net has the capability of generalizing exceptionally well across various datasets [36]. In the ResUNet++ architecture, we have stacked the SE block together with the residual block for improving the performance of the network, increasing the effective generalization across different medical datasets.

C. Atrous Spatial Pyramidal Pooling

Since the introduction of atrous convolution by Chen *et al.* [44] to control the field-of-view to capture contextual information at multi-scale precisely, it has shown promising results for semantic image segmentation. Later, Chen *et al.* [45] proposed ASPP, which is a parallel atrous convolution block to capture multiple-scale information simultaneously. ASPP captures the contextual information at different scales, and multiple parallel atrous convolutions with varying rates in the input feature map are fused [45]. In ResUNet++, we use ASPP as a bridge between the encoder and the decoder sections, and after the final decoder block. We adopt ASPP in ResUNet++ to capture the useful multi-scale information between the encoder and the decoder.

D. Attention Units

Chen *et al.* [46] proposed an attention model that can segment natural images by multi-scale input processing. Attention model is an improvement over average and max-pooling baseline and allows to visualize the features importance at different scales and positions [46]. With the success of attention mechanisms, various medical image segmentation methods have integrated an attention mechanism into their architecture [1], [47]–[49]. The attention block gives importance to the subset of the network to highlight the most relevant information. We conjecture that the attention mechanism in our architecture will boost the effectiveness of the feature maps of the network by capturing the relevant semantic class and filtering out irrelevant information. Motivated by the recent achievements of attention mechanisms in the field of medical image segmentation and computer vision in general, we have integrated an attention block at the decoder part of the ResUNet++ model.

E. Conditional Random Field

Conditional Random Field (CRF) is a popular statistical modeling method used when the class labels for different inputs are not independent (e.g., image segmentation tasks). CRF can model useful geometric characteristics like shape, region connectivity, and contextual information [50]. Therefore, the use of CRF can further improve the models capability to capture contextual information of the polyps and thus improve overall results. We have used CRF as a further step to produce more refined output to the test dataset for improving the segmentation results. We have used a dense CRF for our experiments.

F. Test Time Augmentation

Test-Time Augmentation (TTA) is a technique of performing reasonable modifications to the test dataset to improve the overall prediction performance. In TTA, augmentation is applied to each test image, and multiple augmented images are created. After that, we make predictions on these augmented images, and the average prediction of each augmented image is taken as the final output prediction. Inspired by the improvement of recent SOTA [22], we have used TTA in our work. In this paper, we utilize both horizontal and vertical flip for TTA.

Dataset	Images	Input size	Availability
Kvasir-SEG [4]	1000	Variable	Public
CVC-ClinicDB [11]	612	384×288	Public
CVC-ColonDB [12]	380	574×500	Public
ETIS Larib Polyp DB [13]	196	1225×966	Public
CVC-VideoClinicDB [15], [16][†][◦]	11,954	384×288	Public
ASU-Mayo Clinic Colonoscopy Video Database [14][†]	18,781	688×550	Copyrighted
Kvasir-Sessile[•]	196	Variable	Public

[†]Ground truth for test data not available [◦]Ground truth oval or circle shaped
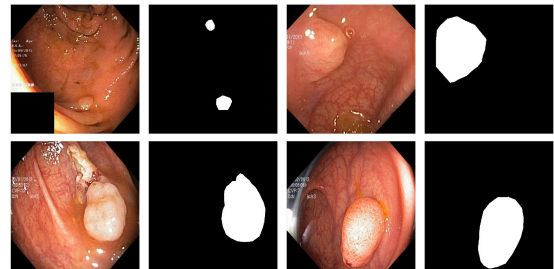[•]Part of Kvasir-SEG [4], only sessile polyps

Fig. 3.   Example polyp and corresponding ground truth from the Kvasir-SEG.

IV. EXPERIMENTS

A. Datasets

We have used six different datasets of segmented polyps with ground truths in our experiments as shown in Table I, i.e., Kvasir-SEG [4], CVC-ClinicDB [11], CVC-ColonDB [12], ETIS Larib Polyp DB [13], CVC-VideoClinicDB [15], [16] and ASU-Mayo Clinic Colonoscopy Video Database [14]. They vary e.g., regarding number of images, image resolution, availability, devices used for capturing and the accuracy of the segmentation masks. One example is given from the Kvasir-SEG in Fig. 3. The Kvasir-SEG dataset includes 196 polyps smaller than 10 mm classified as Paris class 1 sessile or Paris class IIa. We have released this dataset seperately as subset of Kvasir-SEG. Note that for CVC-VideoClinicDB, we have only used the data from the CVC-VideoClinicDBtrainvalid folder since only these data have ground truth masks. Moreover, the ASU-Mayo Clinic Colonoscopy Video Database, which was made available at the "Automatic Polyp Detection in Colonoscopy Videos" sub-challenge at Endovis 2015 had ten normal videos (negative shots) and ten videos with polyps. However, the test subset is not available because of issues related to licensing. In our experiments, while training, validating, testing with 80:10:10 split on the ASU-Mayo, we used all 20 videos for experimentation. However, for the cross-dataset test (i.e., Tables X and XI), we only tested on ten positive polyp videos.

B. Evaluation Method

To evaluate polyp segmentation methods, where individual pixels should be identified and marked, we use metrics used

in earlier research [4], [18], [20], [22], [26], [51] and in competitions like GIANA,[1] comparing the correctly and wrongly identified pixels of findings. The Dice coefficient (DSC) and the Intersection over Union (IoU) are the most commonly used metrics. We use the DSC to compare the similarity between the produced segmentation results and the original ground truth. Similarly, the IoU is used to compare the overlap between the output mask and original ground truth mask of the polyp. The mean Intersection over Union (mIoU) calculates IoU of each semantic class of the image and computes the mean over all the classes. There is a correlation between DSC and mIoU. However, we calculate both the metrics to provide a comprehensive results analysis that could lead to better understanding of the results.

Moreover, other often-used metrics for the binary classification are recall (true positive rate) and precision (positive predictive value). For the polyp segmentation, precision is the ratio of the number of correctly segmented pixels versus the total number of all the pixels. Similarly, recall is the ratio of correctly segmented pixel versus the total number of pixels present in the ground truth. In the polyp image segmentation, precision and recall are used to indicate over-segmentation and under-segmentation. For formal definitions and formulas, see the definitions in for example [4], [51]. Finally, the receiver operating characteristic (ROC) curve analysis is also an important metric to characterize the performance of the binary classification system. In our study, we therefore calculate DSC, mIoU, recall, precision, and ROC when evaluating the segmentation models.

C. Data Augmentation

Data augmentation is a crucial step in increasing the number of polyp samples. This solves the data insufficiency problem, improves the performance of the model, and help to reduce over-fitting. We have used a large number of different data augmentation techniques to increase the training sample. We divide all the polyp datasets into training, validation, and testing sets using the ratio of 80:10:10 based on the random distribution except for the mixed datasets. After splitting the dataset, we apply data augmentation techniques such as center crop, random rotation, transpose, elastic transform, grid distortion, optical distortion, vertical flip, horizontal flip, grayscale, random brightness, random contrast, hue saturation value, RBG shift, course dropout, and different types of blur. For cropping the images, we have used a crop size of 256×256 pixels. For the experiments, we have resized the complete training, validation, and testing dataset to 256×256 pixels to reduce the computational complexity. We have only augmented the training dataset. The validation data is not augmented, and the test datasets were augmented while evaluation using TTA.

D. Implementation and Hardware Details

We have implemented all the models using the Keras framework [52] with Tensorflow [53] as a backend. Source code of our implementation and information about our experimental setup are made publicly available on Github.[2] Our experiments were performed using a Volta 100 Tensor Core GPU on a Nvidia

---

TABLE II
RESULTS COMPARISON ON KVASIR-SEG

Method	DSC	mIoU	Recall	Precision
UNet [35]	0.7147	0.4334	0.6306	**0.9222**
ResUNet [34]	0.5144	0.4364	0.5041	0.7292
ResUNet-mod [34]	0.7909	0.4287	0.6909	0.8713
ResUNet++ [1]	0.8119	0.8068	0.8578	0.7742
ResUNet++ + CRF	0.8129	0.8080	0.8574	0.7775
ResUNet++ TTA	0.8496	0.8318	**0.8760**	0.8203
ResUNet++ +TTA + CRF	**0.8508**	**0.8329**	0.8756	0.8228

DGX-2 AI system capable of 2-petaFLOPS tensor performance. We used an Ubuntu 18.04.3LTS operating system with Cuda 10.1.243 version installed. We have performed different experiments with different sets of hyperparameters manually on the same dataset in order to select the optimal set of hyperparameters for the ResUNet++. Our model performed well with the batch size of 16, Nadam as an optimizer, binary cross-entropy as the loss function, and learning rate of $1e-5$. The dice loss function was also competitive. These hyperparameters were chosen based on the empirical evaluation. All the models were trained for 300 epochs. We have used early stopping to prevent the model from over-fitting. To further improve the results, we have used stochastic gradient descent with warm restarts (SGDR). All the hyperparameters were same except for learning rate, which was adjusted based on the requirement. We have also included the Tensorboard for the analysis and visualization of the results.

V. RESULTS

In our previous work, we have showed that ResUNet++ outperforms the SOTA UNet [35] and ResUNet [34] models trained on Kvasir-SEG and CVC-ClinicDB dataset [1]. In this work, we aim to improve the results of ResUNet++ by utilizing further hyperparameter optimization, CRF and TTA. In this section, we present and compare the results of ResUNet++ with CRF, TTA, and both approaches combined on the same dataset, mixed dataset, and cross-dataset. Although a direct comparison of approaches from the literature is difficult due to different testing mechanisms used by various authors, we nonetheless compare the results with the recent work for the evaluation.

A. Results Comparison on Kvasir-SEG Dataset

Table II and Fig. 4 show the quantitative and qualitative results comparison. Fig. 7 shows the ROC curve for all the models. As seen in the quantitative results (Table II), qualitative results (Fig. 4), and ROC curve (Fig. 7), our proposed methods outperform ResUNet++ on the Kvasir-SEG dataset. The improvement in results demonstrates the advantage of the use of the TTA, CRF and their combinations.

B. Results Comparison on CVC-ClinicDB

CVC-ClinicDB is a commonly used dataset for polyp segmentation. Therefore, it becomes important that we bring different works from the literature together and compare the proposed algorithms with the existing works. We compare our algorithms with the SOTA algorithms. Table III demonstrates that the combination of ResUNet++ and CRF achieves DSC of 0.9293 and mIoU of 0.8898, which is 2.23% improvement on PraNet [57]
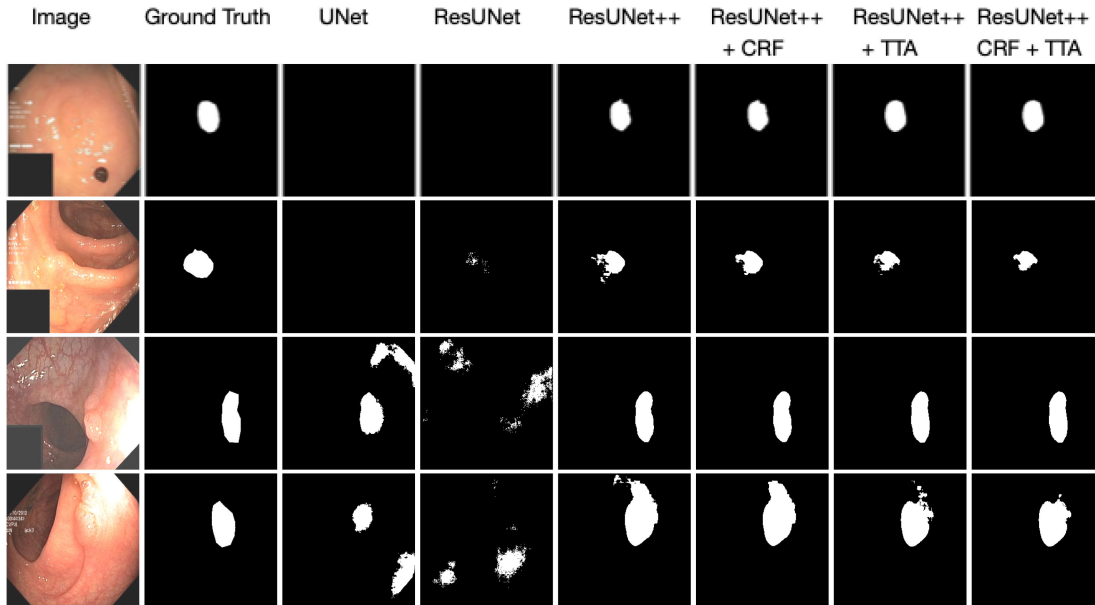
Fig. 4. Qualitative results comparison of the proposed models with UNet, ResUNet, and ResUNet++. The figure shows the example of polyps that are usually missed-out during colonoscopy examination. We see that there is a high similarity between ground truth and predicted mask for the proposed models.

TABLE III
RESULTS COMPARISON ON CVC-CLINICDB

Method	DSC	mIoU	Recall	Precision
MultiResUNet$^\diamond$ [31]	-	0.8497	-	-
cGAN† [24]	0.8848	0.8127	-	-
SegNet [20]	-	-	0.8824	-
FCN$^\bullet$ [54]	-	-	0.7732	**0.8999**
CNN [55]	(0.62-0.87)	-	-	-
MSPB$^\psi$ CNN [56]	0.8130	-	0.7860	0.8090
UNet [35]	0.6419	0.4711	0.6756	0.6868
ResUNet [34]	0.4510	0.4570	0.5775	0.5614
PraNet [57]	0.8980	0.8400	-	-
ResUNet-mod [34]	0.7788	0.4545	0.6683	0.8877
ResUNet++ [1]	0.9199	0.8892	0.9391	0.8445
ResUNet++ + CRF	**0.9203**	**0.8898**	**0.9393**	0.8459
ResUNet++ + TTA	0.9020	0.8826	0.9065	0.8539
ResUNet++ + TTA + CRF	0.9017	0.8828	0.9060	0.8549

†Conditional generative adversarial network $^\diamond$Data augmentation
$^\bullet$Fully convolutional network $^\psi$ multi-scale patch-based

TABLE IV
RESULTS COMPARISON ON CVC-COLONDB

Method	DSC	mIoU	Recall	Precision
FCN-8S + Otsu [19]	0.8100	-	0.7480	-
FCN-8s + Texton [58]	0.7014	-	0.7566	-
SA-DOVA Descriptor [12]	0.5533	-	0.6191	-
PraNet [57]	0.7090	0.6400	-	-
ResUNet++ [1]	0.8469	0.8456	**0.8511**	0.8003
ResUNet++ + CRF	0.8458	0.8456	0.8497	0.7767
ResUNet++ + TTA	**0.8474**	**0.8466**	0.8434	0.8118
ResUNet++ + TTA + CRF	0.8452	0.8459	0.8411	**0.8125**

TABLE V
RESULTS ON ETIS-LARIB POLYP DB

Method	DSC	mIoU	Recall	Precision
PraNet [57]	0.6280	0.5670	-	-
ResUNet++ [1]	**0.6364**	**0.7534**	**0.6346**	0.6467
ResUNet++ + CRF	0.6228	0.7520	0.6242	0.5648
ResUNet++ + TTA	0.6136	0.7458	0.5996	**0.6565**
ResUNet++ + TTA + CRF	0.6018	0.7426	0.5914	0.5755

in DSC and 4.98% improvement in mIoU, respectively, and the proposed methods show the SOTA result on CVC-ClinicDB.

The ROC curve measures the performance for the classification problem provided a set threshold. We have set the probability threshold of 0.5. The combination of ResUNet++ and TTA has the maximum Area Under Curve - Receiver Operating Characteristic (AUC-ROC) of 0.9814, as shown in Fig. 8. Therefore, the results in Table III and Fig. 8 show that applying TTA gives an improvement on CVC-ClinicDB.

C. Results Comparison on CVC-ColonDB Dataset

Our results using the CVC-ColonDB dataset are presented in Table IV. The table shows that the proposed method of combining ResUNet++ and TTA achieved the highest DSC of 0.8474, which is 3.74% higher than SOTA [19], and mIoU of 0.8466 which is 20.66% higher than [57]. The recall and

precision of all three proposed methods are quite acceptable. When compared with ResUNet++, there is an improvement of 1.22% in precision. There are negligible differences in recall, with ResUNet++ slightly outperforming the others.

D. Results Comparison on ETIS-Larib Polyp DB

Table V shows the results of the proposed models on the ETIS-Larib Polyp DB. In this case, we do not compare the results with UNet and ResUNet, but compare the models directly with ResUNet++ as it already showed superior performance on Kvasir-SEG and CVC-ClinicDB [1]. Here, there are only marginal differences in the results of ResUNet++, "ResUNet++ + CRF," "ResUNet++ + TTA," and "ResUNet++ +

TABLE VI
RESULTS ON KVASIR-SESSILE

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.4600	0.64086	0.4382	0.5838
ResUNet++ + CRF	0.4522	0.6394	0.4326	0.5708
ResUNet++ + TTA	**0.5042**	**0.6606**	**0.4851**	**0.6796**
ResUNet++ + TTA + CRF	0.4901	0.6565	0.4766	0.6277

TABLE VII
RESULTS COMPARISON ON CVC-VIDEOCLINICDB

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.8798	0.8730	**0.7749**	0.6702
ResUNet++ + CRF	**0.8811**	**0.8739**	0.7743	**0.6706**
ResUNet++ + TTA	0.8125	0.8467	0.6896	0.6421
ResUNet++ + TTA + CRF	0.8130	0.8477	0.6875	0.6276

TABLE VIII
RESULTS COMPARISON ON ASU-MAYO CLINIC COLONOSCOPY
VIDEO DATABASE

Method	DSC	mIoU	Recall	Precision
ResUNet++ [1]	0.8743	0.8569	**0.6534**	0.4896
ResUNet++ + CRF	**0.8850**	**0.8635**	0.6504	0.4858
ResUNet++ + TTA	0.8553	0.8535	0.6162	**0.4912**
ResUNet++ + TTA + CRF	0.8550	0.8551	0.6107	0.4743

TABLE IX
RESULTS COMPARISON USING (KVASIR-SEG + CVC-CLINICDB) AS THE
TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
CVC-ColonDB	ResUNet++ [1]	0.4974	0.6800	0.4787	**0.6019**
	ResUNet++ + CRF	0.4920	0.6788	0.4744	0.5636
	ResUNet++ + TTA	**0.5084**	**0.6859**	**0.4795**	0.5973
	ResUNet++ + TTA + CRF	0.5061	0.6852	0.4775	0.5770
CVC-Video-ClinicDB	ResUNet++ [1]	0.3460	0.6348	**0.2272**	**0.3383**
	ResUNet++ + CRF	0.3552	0.6412	0.2228	0.3065
	ResUNet++ + TTA	0.3573	0.6440	0.2104	0.3338
	ResUNet++ + TTA + CRF	**0.3603**	**0.6468**	0.2068	0.3038

CRF + TTA". However, ResUNet++ achieves maximum DSC of 0.6364, which is 0.84% improvement over SOTA [57] and mIoU of 0.7534 which is 18.64% improvement over [57]. The recall of ResUNet++ is 0.6346, which is slightly higher than the proposed methods. However, the precision of combining ResUNet++ and TTA is higher as compared to ResUNet++.

From the results, we can conclude that performance of architecture is data specific. Our proposed methods outperformed SOTA over five independent datasets, however, ResUNet++ shows better results than the combinational approaches on the ETIS-Larib dataset. Still, the precision of combining ResUNet++ and TTA is slightly higher than ResUNet++. It is to be noted that ETIS-Larib contains only 196 images, out of which only 156 images are used for training. Even with the small training dataset, the models are performing satisfactorily as compared to the SOTA [57] with significant margin in mIoU, which can be considered as the strength of the algorithm.

E. Results on Kvasir-Sessile

As this is the first work on Kvasir-Sessile, we have compared the proposed methods with ResUNet++. Table VI shows that combining ResUNet++ and TTA gives the DSC of 0.5042, and mIoU of 0.6606, which can be considered a decent score on a smaller size dataset. The dataset contains small, diverse images, which are difficult to generalize with very few training samples.

F. Results Comparison on CVC-VideoClinicDB

Table VII shows the results of the proposed models on the CVC-VideoClinicDB. From the results, we can observe that all models perform well on the dataset despite the fact that masks are not pixel perfect. One of the reasons for high performance is the presence of 11,954 polyps and normal video frames that was used in training and testing. The combination of ResUNet++ and CRF obtained a DSC of 0.8811, mIoU of 0.8739, recall of 0.7743, and precision of 0.6706 which is quite acceptable for the segmentation task with this type of dataset. In CVC-VideoClinicDB, the ground-truth is marked with an oval or circle shape. However, it is understandable that pixel-precise annotations of this dataset will need great manual effort from expert endoscopists and engineers.

G. Results Comparison on AUS-Mayo ClinicDB

Table VIII shows the results of the proposed models on the ASU-Mayo ClinicDB. ASU-Mayo contains 18,781 frames, both polyp and non-polyp images. The combination of ResUNet++ and CRF obtained a DSC of 0.8850 and mIoU of 0.8635. As in the real clinical settings, the models trained on this type of dataset are more meaningful (as it contains both polyp and non-polyp frames). The capability to achieve good performance for these more challenging datasets is one of the strengths of the proposed method. This is supported by the fact that this dataset also contains a sufficient amount of images to enable sufficient training.

H. Results Comparison on Mixed Dataset

To check the performance of the proposed approaches on the images captured using different devices, we have mixed the Kvasir-SEG and CVC-ClinicDB and used them for training. The model were tested on CVC-ColonDB and CVC-VideoClinicDB. Table IX shows the result of the mixed dataset on both datasets. The combination of ResUNet++ and TTA obtains a DSC of 0.5084 and mIoU of 0.6859 with CVC-ColonDB. The combination of ResUNet++, CRF, and TTA obtained a DSC of 0.3603 and mIoU of 0.6468 with CVC-VideoClinicDB.

From the table, we can see that the combination of ResUNet++, CRF, and TTA performs better or very competitive with both still images and video frames. Here, it is also evident that the model trained on the smaller dataset (Kvasir-SEG and CVC-ClinicDB) which do not include non-polyp images, is not performing well on larger and diverse datasets (CVC-VideoClinicDB) that contain both polyp and non-polyp frames. Additionally, for the CVC-VideoClinicDB datasets, the provided ground truth is not perfect (oval/circle) shaped. As the model trained on Kvasir-SEG and CVC-ClinicDB have perfect annotations, the model is good at predicting a perfectly shaped mask. When we make predictions on the CVC-VideoClinicDB with imperfect masks, even if the predictions are good, the scores

TABLE X
CROSS-DATASET RESULTS USING **KVASIR-SEG** AS THE TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
CVC-ClinicDB	ResUNet++ [1]	0.6468	0.7311	0.6984	0.6510
	ResUNet++ + CRF	0.6458	0.7321	0.6955	0.6425
	ResUNet++ + TTA	**0.6737**	**0.7507**	**0.7108**	**0.6833**
	ResUNet++ + TTA + CRF	0.6712	0.7506	0.7078	0.6680
ETIS-Larib Polyp DB	ResUNet++ [1]	**0.4017**	0.6415	**0.4412**	0.3925
	ResUNet++ + CRF	0.4012	0.6427	0.4379	0.3755
	ResUNet++ + TTA	0.4014	**0.6468**	0.4294	**0.4014**
	ResUNet++ + TTA + CRF	0.3997	0.6466	0.4267	0.3710
CVC-ColonDB	ResUNet++ [1]	0.5135	0.6742	0.5398	0.5461
	ResUNet++ + CRF	0.5122	0.6748	0.5367	0.5285
	ResUNet++ + TTA	**0.5593**	**0.7030**	**0.5626**	**0.5944**
	ResUNet++ + TTA + CRF	0.5563	0.7024	0.5595	0.5811
CVC-Video-ClinicDB	ResUNet++ [1]	0.3175	0.6082	**0.2915**	0.3299
	ResUNet++ + CRF	0.3334	0.6185	0.2862	0.3141
	ResUNet++ + TTA	0.3505	0.6337	0.2601	**0.3488**
	ResUNet++ + TTA + CRF	**0.3601**	**0.6402**	0.2555	0.3252
ASU-Mayo	ResUNet++ [1]	0.3482	0.6346	**0.2196**	**0.2021**
	ResUNet++ + CRF	0.3747	0.6516	0.2136	0.1797
	ResUNet++ + TTA	0.3823	0.6583	0.1962	0.2165
	ResUNet++ + TTA + CRF	**0.3950**	**0.6681**	0.1890	0.1781

TABLE XI
CROSS-DATASET RESULTS ON **CVC-CLINICDB** AS THE TRAINING SET

Test set	Method	DSC	mIoU	Recall	Precision
Kvasir-SEG	ResUNet++ [1]	0.6876	0.7374	0.7027	0.7354
	ResUNet++ + CRF	0.6877	0.7389	0.7004	0.7371
	ResUNet++ + TTA	**0.7218**	0.7616	**0.7225**	**0.7855**
	ResUNet++ + TTA + CRF	0.7208	**0.7621**	0.7204	0.7831
CVC-ColonDB	ResUNet++ [1]	0.5489	0.6942	0.5577	0.5816
	ResUNet++ + CRF	0.5470	0.6949	0.5546	0.5727
	ResUNet++ + TTA	**0.5686**	0.7080	**0.5702**	**0.5935**
	ResUNet++ + + TTA + CRF	0.5667	**0.7081**	0.5687	0.5773
ETIS-Larib Polyp DB	FCN-VGG [59]	0.7023	0.5420	-	-
	ResUNet++ [1]	0.4012	0.6398	**0.4232**	0.4013
	ResUNet++ + CRF	0.3990	0.6403	0.4191	0.3974
	ResUNet++ + TTA	0.4027	**0.6522**	0.3969	**0.4235**
	ResUNet++ + TTA + CRF	0.3973	0.6514	0.3906	0.4078
CVC-Video-ClinicDB	ResUNet++ [1]	0.3666	0.6422	**0.2568**	0.3632
	ResUNet++ + CRF	0.3788	0.6500	0.2530	0.3399
	ResUNet++ + TTA	0.3941	0.6582	0.2516	**0.3829**
	ResUNet++ + TTA + CRF	**0.3988**	**0.6616**	0.2481	0.3542
ASU-Mayo	ResUNet++ [1]	0.2797	0.6113	**0.1627**	0.1443
	ResUNet++ + CRF	0.3167	0.6323	0.1591	0.1348
	ResUNet++ + TTA	0.3085	0.6331	0.1265	**0.1571**
	ResUNet++ + TTA + CRF	**0.3233**	**0.6426**	0.1225	0.1270

may not be high because of the difference in the provided ground truth and the predicted masks.

I. Cross-Dataset Result Evaluation on Kvasir-SEG

For the cross-dataset evaluation, we trained the models on the Kvasir-SEG dataset and tested it on the other five independent datasets. Table X shows the results of cross-data generalizability of ResUNet++ alone, and with the CRF and TTA techniques. The results of the models trained on Kvasir-SEG produces an average best mIoU of 0.6817 and an average best DSC of 0.4779 for both image and video datasets. From the above table, we can observe that the proposed combinational approaches are performing competitive. For the image datasets, the combination of ResUNet++ and TTA is performing better, and for the video datasets, the combination of ResUNet++, CRF, and TTA is performing best. It is to be noted that we are training a model with 1000 Kvasir-SEG pixel segmented polyps and testing on (for example, 11,954 frames) oval-shaped polyp ground truth. Here, even if the predictions are correct, the evaluation scores will not be good because of the oval/circle shaped ground truth. Moreover, the datasets such as ASU-Mayo and CVC-VideoClinicDB are heavily imbalanced, but the model trained on Kvasir-SEG contains at least one polyp. This may also have caused the poor performance.

J. Cross-Dataset Evaluation on CVC-ClinicDB

To further test generalizability, we trained the models on CVC-CliniDB and tested it across five independent, diverse image and video datasets. Tables XI shows the results of cross-data generalizability. Like the previous test on Kvasir-SEG, the results follow the same pattern with the combination of ResUNet++ and TTA outperforming others on the image datasets and the combination of ResUNet++, CRF, and TTA outperforming its competitors on video datasets. ResUNet++ and TTA still remain competitive. Moreover, the values of DSC and mIoU of the best model are similar for both the CVC-VideoClinicDB

and the ASU-Mayo Clinic Colonoscopy Video Database. We have compared the results with the existing work that used CVC-CliniDB for training and ETIS-Larib for testing. Our model achieves highest mIoU of 0.6522.

K. Result Summary

In summary, from all obtained results (i.e., qualitative, quantitative, and ROC curve), the following main observations can be drawn: (i) the proposed ResUNet++ is capable of segmenting the smaller, larger, and regular polyps; (ii) the combination of ResUNet++ with CRF achieves the best performance in terms of DSC, mIoU, recall, and precision when trained and tested on the same dataset (see Table III, Table VII, and Table VIII) whereas it remains competitive when tested on other datasets; (iii) the combination of ResUNet++ and TTA and the combination of ResUNet++, CRF, and TTA performs similar for the mixed datasets; (iv) the combination of ResUNet++ and TTA outperforms others on still images; (v) the combination of ResUNet++, CRF and TTA shows improvement on all the video datasets compared to ResUNet++; (vi) all the models perform better when the images have higher contrast; (vii) ResUNet++ is particularly good at segmenting smaller and flat or sessile polyps, which is a prerequisite for developing an ideal CADx polyp detection system [1]; (viii) ResUNet++ fails especially on the images that contains over-exposed regions termed as saturation or contrast (see Fig. 6); (ix) ResUNet and ResUNet-mod particularly showed over-segmented or under-segmented results (see Fig. 4).

VI. DISCUSSION

A. General Performance

The tables and figures suggest that applying CRF and TTA improved the performance of ResUNet++ on the same datasets, mixed dataset, and cross-datasets. Specifically, the combination of ResUNet++ and TTA, and the combination of ResUNet++,

TABLE XII
TOTAL NUMBER OF TRAINABLE PARAMETERS

Model	Trainable parameters
U-Net	5,400,289
ResUNet	8,221,121
ResUNet-mod	2,058,465
ResUNet++	16,228,001

CRF, and TTA are more generalizable for all the datasets, where TTA with ResUNet++ performs best on the still images, and the combinations of ResUNet++, CRF, and TTA are outperforming others on video datasets. For all of the proposed models, the value of AUC is greater than 0.93. This indicates that our models are good at distinguishing between the polyp and non-polyps. It also suggests that the model produces sufficient sensitivity.

The total number of trainable parameters increases by increasing the number of blocks in the networks (see Table XII). However, in ResUNet++, there is significant performance gain that compensates for the training time, and our model requires fewer parameters if we compare with the models that use pre-trained encoders.

B. Cross Dataset Performance

The cross-data test is an excellent technique to determine the generalizing capability of a model. The presented work is an initiative towards improving the generalizability of segmentation methods. Our contribution towards generalizability is to train on one dataset and test on several other public datasets that may come from different centers and use different scope manufacturers. Thus, we conjecture that to tackle this issue, out-of-sample multicenter data must be used to test the built methods. The work is a step forward in raising an issue regarding method interpretability, and we also raise questions about generalizability and domain adaptation of supervised methods in general.

From the results analyses, we can see that different proposed algorithms perform well with different types of datasets. For instance, CRF outperformed others on tables III, VII, and VIII. TTA showed improvement on tables IV, IX, X, and XI. CRF performs better than TTA while trained and tested on video datasets (see tables VII and VIII). CRF also outperformed TTA on most of the images dataset. However, TTA still remains competitive. On the mixed dataset and the cross-dataset test, TTA performs better than CRF on all the datasets. On the mixed datasets and on the cross-dataset test on videos, the combination of ResUNet++, CRF, and TTA remains the best choice (see tables IX, X, and XI). There is a performance improvement over ResUNet++ while combining CRF, TTA, and the combination of CRF and TTA.

However, there is no significant performance improvement of any methods over the others. From the results, we can observe that the results are typically data-dependent. However, as the proposed methods perform well on video frames, it may work better in the clinic, as the output from a colonoscope is a video stream. Thus, it becomes critical to show the results with all three approaches on each dataset. Therefore, we provide extensive
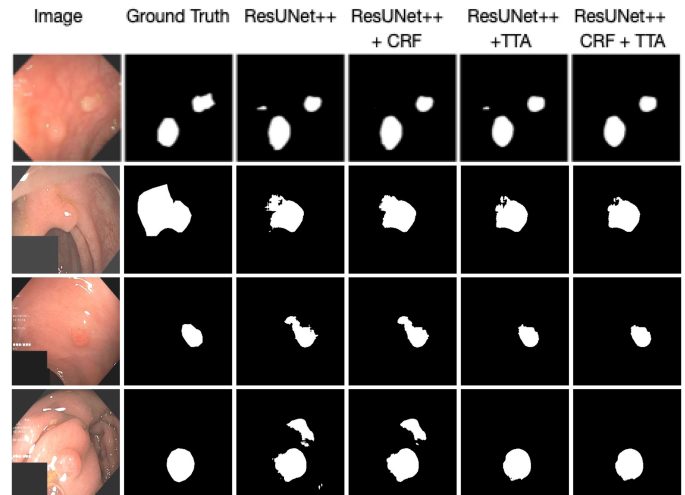

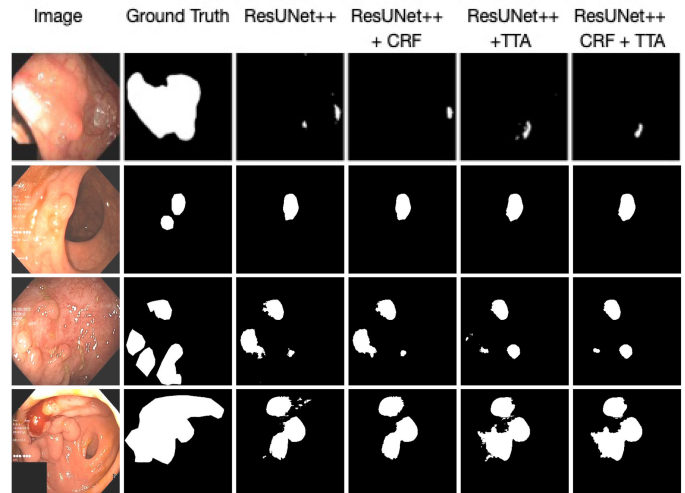Fig. 5. Result of model trained on CVC-ClinicDB and tested on Kvasir-SEG.

Fig. 6. Example images where the proposed models fails on Kvasir-SEG.

experiments showing both success (Fig. 4, Fig. 5) and failure cases (Fig. 6) and present the overall analysis.

C. Challenges

There are several challenges associated with segmenting polyps, such as bowel-quality preparation during colonoscopy, angle of the cameras, superfluous information, and varying morphology, which can affect the overall performance of a DL model. For some of the images, there even exists variation in the decision between endoscopists. While ResUNet++ with CRF and TTA also struggle with producing satisfactory segmentation maps for these images, it performs considerably better than our previous model and also outperforms another SOTA algorithm.

The quality of a colonoscopy examination is largely determined by the experience and skill of the endoscopist [23]. Our proposed model can help in two ways: (i) it can be used to
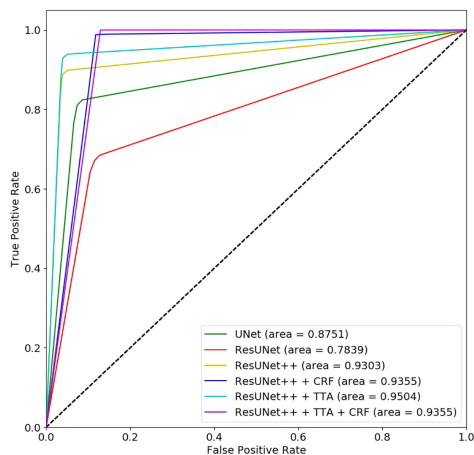
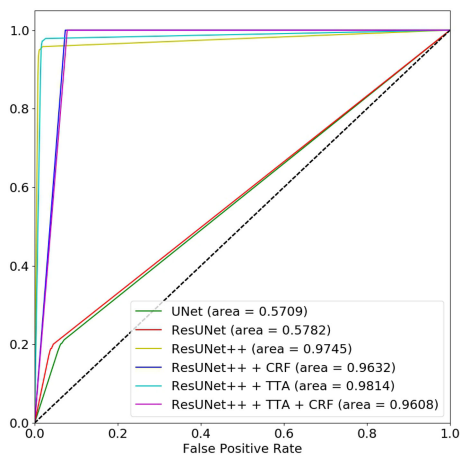Fig. 7. ROC curve of proposed models on the Kvasir-SEG.

Fig. 8. ROC curve for all the models trained and tested on CVC-ClinicDB.

segment a detected polyp, providing an extra pair of eyes to the endoscopist; and (ii) it performs well on both flat and small polyps, which are often missed during endoscopic examinations. The qualitative analysis (see Fig. 4) and the quantitative analyses from the above tables and figures support this argument. This is a major strength of our work and makes it a candidate for clinical testing.

D. Possible Limitations

The retrospective design is a possible limitation of this study and prospective studies are essential because less prone to bias. Prospective clinical evaluation is essential because data analyzed with the retrospective study is different prospective study (for example, the case of missing data that should be considered on the basis of best-case and worse case scenarios) [60]. Also, all data in these experiments are curated, while a prospective clinical trial would mean testing on full colonoscopy videos. During model training, we have resized all the images to 256×256 to reduce the complexity, which costs in loss of information, and can affect the overall performance. We have worked on optimizing the code, but further optimization may exist, that can potentially improve the performance of the model.

VII. CONCLUSION

In this paper, we have presented the ResUNet++ architecture for semantic polyp segmentation. We were inspired from the residual block, ASPP, and attention block to design the novel ResUNet++ architecture. Furthermore, we applied CRF and TTA to improve the results even more. We have trained and validated the combination of ResUNet++ with CRF and TTA using six publicly available datasets, and analyzed and compared the results with the SOTA algorithm on specific datasets. Moreover, we analyzed the cross-data generalizability of the proposed model towards developing generalizable semantic segmentation models for automatic polyp segmentation. A comprehensive evaluation of the proposed model trained and tested on six different datasets showed good performance of the (ResUNet++ and CRF) on image datasets and (ResUNet++ and TTA), (ResUNet++, CRF, and TTA) model for the mixed datasets and cross-datasets. Further, a detailed study on cross-dataset generalizability of the models trained on Kvasir-SEG and CVC-ClinicDB and tested on five independent datasets, confirmed the robustness of the proposed ResUNet++ + TTA method for cross-dataset evaluation.

The strength of our method is that we successfully detected smaller and flat polyps, which are frequently missed during colonoscopy examination [20], [61]. Our model can also detect the polyps that would be difficult for the endoscopists to identify without careful investigations. Therefore, we conjecture that the ResUNet++ architecture, along with the additional CRF and TTA steps, could be one of the potential areas to investigate further, especially for the overlooked polyps. We also point out that the lack of generalization issues of the models, which is evidenced by the unsatisfactory result for cross-dataset evaluation in most of the cases. In the future, our CADx system should also be investigated on other bowel conditions. Moreover, a prospective clinical trial needs to be conducted to demonstrate the usefulness of the proposed system.

REFERENCES

[1] D. Jha *et al.*, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia.*, 2019, pp. 225–230.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[3] T. Matsuda, A. Ono, M. Sekiguchi, T. Fujii, and Y. Saito, "Advances in image enhancement in colonoscopy for detection of adenomas," *Nature Rev. Gastroenterol. Hepatol.*, vol. 14, no. 5, pp. 305–314, 2017.

[4] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Model.*, 2020, pp. 451–462.

[5] S. B. Ahn, D. S. Han, J. H. Bae, T. J. Byun, J. P. Kim, and C. S. Eun, "The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies," *Gut Liver*, vol. 6, no. 1, pp. 64–70, 2012.

[6] D. o. Heresbach, "Miss rate for colorectal neoplastic polyps: A prospective multicenter study of back-to-back video colonoscopies," *Endoscopy*, vol. 40, no. 4, pp. 284–290, 2008.

[7] Zimmermann-Fraedrich *et al.*, "Right-sided location not associated with missed colorectal adenomas in an individual-level reanalysis of tandem colonoscopy studies," *Gastroenterol.*, vol. 157, no. 3, pp. 660–671, 2019.

[8] A. Shaukat *et al.*, "Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy," *Gastroenterol.*, vol. 149, no. 4, pp. 952–957, 2015.

[9] D. Vàzquez *et al.*, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, 2017.

[10] T. Roß *et al.*, "Robust medical instrument segmentation challenge 2019," 2020, *arXiv:2003.10299v1.*

[11] J. Bernal, F. J. Sànchez, G. Fernàndez-Esparrach, D. Gil, C. Rodrìguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[12] J. Bernal, J. Sànchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[13] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *Int. J. Comput. Assis. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[14] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2015.

[15] Q. Angermann *et al.*, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Comput. Assis. Robot. Endos. Clin. Image-Based Proced.*, 2017, pp. 29–41.

[16] J. Bernal *et al.*, "Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases," in *Proc. CARS Conf.*, 2018.

[17] A. A. Pozdeev, N. A. Obukhova, and A. A. Motyko, "Automatic analysis of endoscopic images for polyps detection and segmentation," in *Proc. Conf. Russian Young Res. Elect. Electron. Eng.*, 2019, pp. 1216–1220.

[18] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[19] M. Akbari *et al.*, "Polyp segmentation in colonoscopy images using fully convolutional network," in *Proc. Int. Eng. Med. Biol. Conf.*, 2018, pp. 69–72.

[20] P. Wang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, 2018.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[22] Y. B. Guo and B. Matuszewski, "Giana polyp segmentation with fully convolutional dilation neural networks," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2019, pp. 632–641.

[23] M. Yamada *et al.*, "Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.

[24] J. Poomeshwaran, K. S. Santhosh, K. Ram, J. Joseph, and M. Sivaprakasam, "Polyp segmentation using generative adversarial network," in *Proc. Int. Eng. Med. Biol. Conf.*, 2019, pp. 7201–7204.

[25] J. Kang and J. Gwak, "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images," *IEEE Access*, vol. 7, pp. 26 440–26 447, 2019.

[26] S. Ali *et al.*, "Endoscopy artifact detection (EAD 2019) challenge dataset," 2019, *arXiv:1905.03209.*

[27] N.-Q. Nguyen and S.-W. Lee, "Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network," *IEEE Access*, vol. 7, pp. 33 795–33 808, 2019.

[28] V. de Almeida Thomaz, C. A. Sierra-Franco, and A. B. Raposo, "Training data enhancements for robust polyp segmentation in colonoscopy images," in *Proc. Int. Symp. Comput.-Based Med. Syst.*, 2019, pp. 192–197.

[29] X. Sun, P. Zhang, D. Wang, Y. Cao, and B. Liu, "Colorectal polyp segmentation by u-net with dilation convolution," in *Proc. IEEE Int. Conf. Mach. Learn. App.*, 2019, pp. 851–858.

[30] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, 2020, pp. 558–564.

[31] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[32] P. Brandao *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks," *J. Med. Robot. Res.*, vol. 3, no. 2, 2018, Art. no. 1840002.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[34] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geosci. Remo. Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[37] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587.*

[38] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. Rese.*, 2015, pp. 448–456.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] L. Wang, R. Chen, S. Wang, N. Zeng, X. Huang, and C. Liu, "Nested dilation network (NDN) for multi-task medical image segmentation," *IEEE Access*, vol. 7, pp. 44 676–44 685, 2019.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062.*

[45] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[46] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[47] Y. Wang *et al.*, "Deep attentional features for prostate segmentation in ultrasound," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 523–530.

[48] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 370–378.

[49] A. Sinha and J. Dolz, "Multi-scale guided attention for medical image segmentation," *IEEE J. Biomed Health Inform.*, vol. 25, no. 1, pp. 121–130, 2021.

[50] F. I. Alam *et al.*, "Conditional random field and deep feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remo. Sens.*, vol. 57, no. 3, pp. 1612–1628, Mar. 2019.

[51] K. Pogorelov *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. Multimedia Syst. Online Conf.*, 2017, pp. 164–169.

[52] F. Chollet *et al.*, "Keras," 2015.

[53] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. Ocular Surf. Dis. Inde*, 2016, pp. 265–283.

[54] Q. Li *et al.*, "Colorectal polyp segmentation using a fully convolutional neural network," in *Proc. Int. Congr. Image Signal Process., BioMed. Eng. Inform.*, 2017, pp. 1–5.

[55] Q. Nguyen and S.-W. Lee, "Colorectal segmentation using multiple encoder-decoder network in colonoscopy images," in *Proc. Conf. Inf. Knowl. Eng.*, 2018, pp. 208–211.

[56] D. Banik, D. Bhattacharjee, and M. Nasipuri, "A multi-scale patch-based deep learning system for polyp segmentation," *Advan. Comput. Syst. Secur.*, 2020, pp. 109–119.

[57] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. of MICCAI*, 2020, pp. 263–273.

[58] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. Conf. Med. Image Understanding Anal.*, 2017, pp. 707–717.

[59] P. Brandao *et al.*, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in *Med. Imag.*: *Comput.-Aided Diagnosis*, vol. 10134, 2017, pp. 101 340F1–101 340F1.

[60] Y. Mori and S.-e. Kudo, "Detecting colorectal polyps via machine learning," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 713–714, 2018.

[61] A. Leufkens, M. Van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.