Machine Intelligence Research
www.mi-research.net

19(6), December 2022, 531-549
DOI: 10.1007/s11633-022-1371-y

Video Polyp Segmentation: A Deep Learning Perspective

Ge-Peng Ji[1†] Guobao Xiao[2†] Yu-Cheng Chou[3†] Deng-Ping Fan[4*]
Kai Zhao[5] Geng Chen[6] Luc Van Gool[4]

[1] Research School of Engineering, Australian National University, Canberra 2601, Australia
[2] College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China
[3] Department of Computer Science, Johns Hopkins University, Baltimore 21218, USA
[4] Computer Vision Laboratory, ETH Zürich, Zürich 8092, Switzerland
[5] Department of Radiological Sciences, University of California, Los Angeles 90095, USA
[6] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi′an 710072, China

Abstract: We present the first comprehensive video polyp segmentation (VPS) study in the deep learning era. Over the years, developments in VPS are not moving forward with ease due to the lack of a large-scale dataset with fine-grained segmentation annotations. To address this issue, we first introduce a high-quality frame-by-frame annotated VPS dataset, named SUN-SEG, which contains 158 690 colonoscopy video frames from the well-known SUN-database. We provide additional annotation covering diverse types, i.e., attribute, object mask, boundary, scribble, and polygon. Second, we design a simple but efficient baseline, named PNS+, which consists of a global encoder, a local encoder, and normalized self-attention (NS) blocks. The global and local encoders receive an anchor frame and multiple successive frames to extract long-term and short-term spatial-temporal representations, which are then progressively refined by two NS blocks. Extensive experiments show that PNS+ achieves the best performance and real-time inference speed (170 fps), making it a promising solution for the VPS task. Third, we extensively evaluate 13 representative polyp/object segmentation models on our SUN-SEG dataset and provide attribute-based comparisons. Finally, we discuss several open issues and suggest possible research directions for the VPS community. Our project and dataset are publicly available at https://github.com/GewelsJI/VPS.

Keywords: Video polyp segmentation (VPS), dataset, self-attention, colonoscopy, abdomen.

Citation: G. P. Ji, G. Xiao, Y. C. Chou, D. P. Fan, K. Zhao, G. Chen, L. V. Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, vol.19, no.6, pp.531–549, 2022. http://doi.org/10.1007/s11633-022-1371-y

1 Introduction

As the second most deadly cancer and the third most common malignancy, colorectal cancer (CRC) is estimated to cause millions of incidence cases and deaths yearly. The survival rate of CRC patients is higher than 95% in the first stage of the disease but, dramatically decreases to less than 35% in the fourth and fifth stages[1]. Therefore, the early diagnosis of positive CRC cases through screening techniques, such as colonoscopy and sigmoidoscopy, is vital in increasing the survival rate. For prevention purposes, physicians can remove the colon polyps that are at risk of turning into cancer. However, this process highly depends on the physicians′ experience and suffers from a high polyp missing rate, i.e., 22%–28%[2].

Recently, artificial intelligence (AI) techniques have been applied to the automatic detection of candidate lesion polyps during colonoscopy for physicians. However, developing AI models with a satisfactory detection rate is still challenging due to two problems: 1) Limited annotated data. Deep learning models are often hungry for a large-scale video dataset with densely-annotated labels. Moreover, a community-agreed benchmark is missing for evaluating the approaches′ actual performance. 2) Dynamic complexity. The colonoscopy usually involves less ideal conditions of camera-moving acquisition, such as the diversity of colon polyps (e.g., boundary contrast, shape, orientation, shooting angle), internal artifacts (e.g., water flow, residue), and imaging degradation (e.g., color distortion, specular reflection). To this end, we present a systematic study to facilitate the development of deep learning models for video polyp segmentation (VPS). The main contributions of this work are summarized as following three points:

VPS dataset. We elaborately introduce a large-scale VPS dataset, termed SUN-SEG, containing 158 690 frames selected from the SUN-database[3]. We provide a variety of labels, including attribute, object mask, bound-

Research Article
Manuscript received July 3, 2022; accepted August 24, 2022; published online November 3, 2022
Recommended by Associate Editor Jordi Gonzalez

* Corresponding author is Deng-Ping Fan (E-mail: dengpfan@gmail.com)

† These authors contribute equally to this work

ary, scribble, and polygon. These labels can further support the development of colonoscopy diagnosis, localization, and derivative tasks.

VPS baseline. We design a simple but efficient VPS baseline, named PNS+, which consists of a global encoder, a local encoder, and two normalized self-attention (NS) blocks. The global and local encoders extract long- and short-term spatial-temporal representations from the first anchor frame and multiple successive frames, respectively. The NS block dynamically updates the receptive field when coupling attentive cues among extracted features. Experiments show that PNS+ achieves the best performance on the challenging SUN-SEG dataset.

VPS benchmark. To comprehensively understand VPS development, we conduct the first large-scale benchmark by evaluating 13 cutting-edge polyp/object segmentation approaches. Based on the benchmarking results (i.e., five image-based and eight video-based), we argue that the VPS task is not well undertaken and leaves plenty of room for further exploration.

A preliminary version of this work was presented in [4]. In this extended work, we introduce three different contributions. In Section 3, we introduce a high-quality densely-annotated VPS dataset, SUN-SEG, with five extended labels, i.e., attribute, object mask, boundary, scribble, and polygon. Based on the normalized self-attention block as in [4], we propose a global-to-local learning paradigm to realize the modeling of both long-term and short-term dependencies. This part is detailed in Section 4.3. As shown in Section 5, we construct the first large-scale VPS benchmark, which contains 13 of the latest polyp/object segmentation competitors. We highlight several potential research directions based on the above benchmark results and progress in the VPS field.

2 Related works

This section reviews the recent efforts in computer-aided polyp diagnosis from the following two aspects: colonoscopy-related datasets (Section 2.1) and approaches (Section 2.2).

2.1 Colonoscopy-related datasets

Several datasets have been collected for the examination of human colonoscopy. As shown in Table 1, we summarize some key statistics from 20 popular datasets and our SUN-SEG dataset. In light of the task definition,

Table 1 Statistics of existing 20 datasets for human colonoscopy. #IMG = Number of images; #VID = Number of video sequences; DL = Densely labeling; CLS = Classification label; BBX = Bounding box; PM = Pixel-level mask.

Dataset	Year	#IMG	#VID	DL	CLS	BBX	PM	Website
CVC-ColonDB[1]	2012	300	13				✓	–
ETIS-Larib[5]	2014	196	34				✓	–
CVC-ClinicDB[6]	2015	612	31				✓	Link
ColonoscopicDS[7]	2016	–	76		✓			Link
ASU-Mayo[8]	2016	36 458	38	✓			✓	Link
CVC-ClinicVideoDB[9]	2017	11 954	18	✓		✓		Link
CVC-EndoSceneStill[10]	2017	912	44				✓	–
KID2[11, 12]	2017	2 371	47		✓		✓	Link
Kvasir[13]	2017	8 000	–		✓			Link
EDD2020[14]	2020	386	–		✓	✓	✓	Link
SUN-database[3]	2020	158 690	113	✓	✓	✓		Link
Hyper-Kvasir[15]	2020	110 079	374		✓	✓	✓	Link
Kvasir-SEG[16]	2020	1 000	–				✓	Link
PICCOLO[17]	2020	3 433	40		✓		✓	Link
Kvasir-Capsule[18]	2021	4 741 504	117	✓	✓	✓		Link
CP-CHILD-A[19]	2021	8 000	–		✓			Link
CP-CHILD-B[19]	2021	1 500	–		✓			Link
LDPolypVideo[20]	2021	40 266	160	✓	✓	✓		Link
KUMC[21]	2021	37 899	155	✓	✓	✓		Link
PolypGen[22]	2021	6 282	26		✓	✓	✓	Link
SUN-SEG (our)	2022	158 690	1 013	✓	✓	✓	✓	Link

we categorize them into three main-stream partitions.

2.1.1 Classification

There are four popular datasets for the initial purpose of identifying gastrointestinal lesions. ColonoscopicDS[7] collects 76 regular colonoscopy videos with three types of gastrointestinal lesions, including hyperplasic, serrated, and adenoma lesions. Kvasir[13] contains eight types of anatomical landmarks (i.e., polyps, esophagitis, ulcerative colitis, z-line, pylorus, cecum, dyed polyps, and dyed resection margins), and each type has 1 000 images. Hyper-Kvasir[15] further collects 110 079 samples from 374 colonoscopy videos, containing three types of annotations: 10 662 class labels with 23 different lesion findings and 1 000 images with segmented masks and bounding box labels. Notably, all the segmented masks in Hyper-Kvasir are selected from Kvasir-SEG[15]. Recently, CP-CHILD-A & CP-CHILD-B[19] record the colonoscopy data from children, including two classes (i.e., colon polyp, normal or other pathological images) for the classification task.

2.1.2 Detection

There are five widely-accepted video datasets mainly used for the detection task. CVC-ClinicVideoDB[9], as the early video dataset, comprises 18 videos with a total number of 11 954 frames, of which 10 025 frames contain at least a polyp. As for the largest densely-annotated video polyp detection dataset, the SUN-database[3] consists of 49 136 positive samples with their bounding boxes acquired from 99 patients. More recently, two video datasets (i.e., Kvasir-Capsule[18] and KUMC[21]) have been applied for both detection and classification tasks. Especially, the former provides 47 238 bounding box labels from 14 lesion classes, and the latter has 37 899 frames with bounding box labels. Unlike the above datasets, LD-PolypVideo[20] includes 40 266 frames with circular annotations from 160 colonoscopy videos.

2.1.3 Segmentation

As for the video datasets, the early benchmark CVC-EndoSceneStill[10] opts for the combination of CVC-ColonDB[1] and CVC-ClinicDB[6]. ETIS-Larib[5] provides 196 labeled samples from 32 colonoscopy videos, containing about five frames for each sequence. EDD2020[14] contains 386 endoscopy images from five different institutions and multiple gastrointestinal organs. They provide annotations for disease detection, localization, and segmentation. PICCOLO[17] also samples 3 433 frames from 40 videos with sparse annotations. As such, the above five video datasets adopt the sampling annotation strategy, which still lacks per-frame masks on each video sequence due to the labor-intensive annotation process. Being the pioneering video dataset with densely-annotated masks, ASU-Mayo[8] contains 36 458 continuous frames from 38 videos, while it only provides 3 856 labels for ten positive videos. Recently, PolypGen[22] has collected a multi-center dataset incorporating more than 300 patients, including single and continuous frames with 3 788 annotated segmentation masks and bounding box labels. Unlike existing works, we introduce SUN-SEG, the first high-quality densely-annotated dataset for the VPS task, which contains rich annotated labels, such as object mask, boundary, scribble, polygon, and attribute. We hope that this work could fuel the development of colonoscopy diagnosis, localization, and derivative tasks.

2.2 Colonoscopy-related methods

Early solutions[1, 23–25] have been dedicated to identifying colon polyps via mining hand-crafted patterns, such as color, shape, texture, and super-pixel. However, they usually suffer from low accuracy due to the limited capability of representing heterogeneous polyps, as well as the close resemblance between polyps and hard mimics[26]. In contrast, data-driven AI techniques can handle these challenging conditions with better learning ability. This section mainly focuses on tracking the latest image/video polyp segmentation techniques[27], while leaving the systematic review of polyp classification[28, 29] and detection[30, 31] in our future work.

2.2.1 Image polyp segmentation (IPS)

Several methods have been proposed to locate the pixel-level polyp regions from the colonoscopy images. They can be grouped into two major categories. 1) CNN-based approaches. Brandao et al.[32] adopted a fully convolutional network (FCN) with a pre-trained model to segment polyps. Later, Akbari et al.[33] introduced a modified FCN to improve the segmentation accuracy. Inspired by the vast success of UNet[34] in biomedical image segmentation, UNet++[35] and ResUNet[36] were employed for polyp segmentation for improved performance. Furthermore, PolypSeg[37], ACS[38], ColonSegNet[39], and SCR-Net[40] explore the effectiveness of UNet-enhanced architecture on adaptively learning semantic contexts. As the newly-proposed methods, SANet[41] and MSNet[42] design the shallow attention module and subtraction unit, respectively, to achieve precise and efficient segmentation. Additionally, several works opt for introducing additional constraints via three main-stream manners: exerting explicit boundary supervision[43–47], introducing implicit boundary-aware representation[48–50], and exploring uncertainty for ambiguous regions[51]. 2) Transformer-based approaches. Recently, Transformers[52] have been gaining popularity thanks to their powerful modeling ability. TransFuse[53] combines the Transformer and CNN, termed the parallel-in-branch scheme, for capturing global dependencies and low-level spatial details. Besides, a BiFusion module was designed to fuse multi-level features from both branches. Segtran[54] proposes a squeezed attention block that regularizes self-attention, and the expansion block learns diversified representations. A positional encoding scheme was proposed to impose an inductive continuity bias. Based on PVT[55], Dong et al.[56] introduced a model with three tight components, i.e., cascaded fusion, camouflage identification, and similarity aggregation modules.

2.2.2 Video polyp segmentation (VPS)

Despite their progress, existing IPS methods suffer from an inherent limitation of overlooking the valuable temporal cues in colonoscopy videos. Therefore, efforts have been dedicated to combining spatial-temporal features among consecutive video frames. A hybrid 2/3D CNN framework[2] was proposed to aggregate spatial-temporal correlations and achieved better segmentation results. However, the kernel size restricts the spatial correlation between frames, restricting the accurate segmentation of the fast movements of polyps. To alleviate the above problem, PNSNet[4] introduces a normalized self-attention (NS) block to learn spatial-temporal representations with neighborhood correlations effectively. In this paper, we delve deeper into a more effective global-to-local learning strategy based on the NS block, which can fully leverage both long-term and short-term spatial-temporal dependencies.

3 VPS dataset

We describe the introduced SUN-SEG dataset′s details in terms of data collection/re-organization (Section 3.1), professional annotations (Section 3.2), and dataset statistics (Section 3.3).

3.1 Data organization

The colonoscopy videos in our SUN-SEG are from the Showa University and Nagoya University databases (also named SUN-database[3]), the largest video polyp dataset for the detection task. There are two advantages of adopting the SUN-database as our data source. 1) Challenging scenarios. The videos are captured by the high-definition endoscope (CF-HQ290ZI & CF-H290ECI, Olympus) and video recorder (IMH-10, Olympus), providing videos of various polyp sizes in dynamic scenarios, such as imaging at different focusing distances and speeds. 2) Reliable pathological localization. The initial classification information and bounding box annotations are provided by three research assistants and examined by two expert endoscopists with professional domain knowledge.

The original SUN-database has 113 colonoscopy videos, including 100 positive cases with 49 136 polyp frames and 13 negative cases with 109 554 non-polyp frames[1]. We manually trim them into 378 positive and 728 negative clips while maintaining their consecutive intrinsic relationship. Such data pre-processing ensures that each clip has around 3–11 s duration at a real-time frame rate (i.e., 30 fps), promoting the fault-tolerant margin for various algorithms and devices. To this end, the re-organized SUN-SEG contains 1 106 short video clips with 158 690 video frames in total, offering a solid foundation to build such a representative benchmark.

3.2 Professional annotations

Following [57], we adopt a similar annotation pipeline. According to the origin bounding box labels of the SUN-database[3], ten experienced annotators are instructed to offer various labels using Adobe Photoshop. Then, three colonoscopy-related researchers re-verify the quality and correctness of these initial annotations. Fig. 1 shows two typical samples under the restricted quality controls (i.e., rejected and passed). In addition to the original pathological materials provided by SUN-database, such as pathological pattern (e.g., low-grade adenoma, hyperplastic polyp, etc.), shape (e.g., pedunculated, subpedunculated, etc.), and location (e.g., cecum, ascending colon, etc.), we further extend them with diversified annotations in our SUN-SEG. The newly-extended annotations consist of the following five hierarchies: visual attribute → object mask → boundary → scribble → polygon. Selected samples and corresponding annotations can be found in Fig. 2 and their illustrations[2] are as follows.

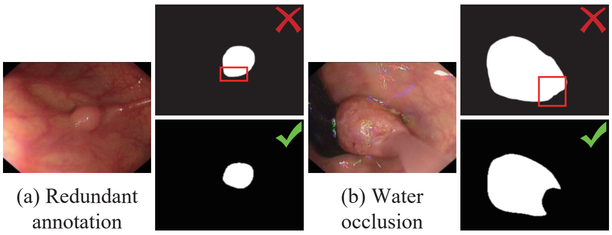


Fig. 1 High-criteria control for data annotation. For instance, we reject case (a), where the boundary is not consistent with the polyp, and case (b), where the water overlapping area is falsely annotated.

Visual attribute. According to the visual characteristics of the videos, we provide ten visual attributes at the video level, whose classification criteria are detailed in Table 2.

Object mask. Correctly parsing lesion areas is helpful for a clinician. Therefore, as shown in Fig. 2(a), we provide pixel-wise object masks for each frame. We further refine the coordinates of the original bounding box based on the object mask to tighten the target, offering more reliable localization labels.

[1] These statistical data come from this website, http://amed8k.sundatabase.org/, which is different from the data reported in the original paper[3]. Besides, the SUN-database is available for only non-commercial use in research or educational purposes, which could be freely accessed with permission from the authors.

[2] The descriptions of complete annotations refer to https://github.com/GewelsJI/VPS/blob/main/docs/DATA_DESCRIPTION.md.

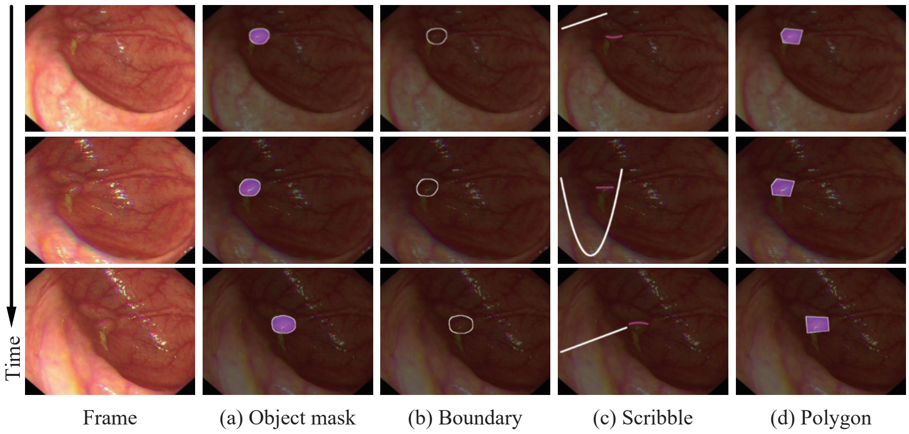


Fig. 2 Diversified annotations for each video frame in our SUN-SEG dataset, including object mask (a), boundary (b), and two weak labels, i.e., scribble (c) and polygon (d). For more details, refer to Section 3.2.

Boundary. Fig. 2(b) shows the polyp boundary generated by calculating the gradient of the object mask.

Scribble. Besides, we offer two weak labels to facilitate the research under data-insufficient conditions. As for the scribble labels in Fig. 2(c), we use two high-degree curves to indicate the foreground (purple curve) and background (white curve), respectively. To ensure the objectivity of various annotators, we adopt linear or quadratic functions to randomly create the above curves in the positive/negative region.

Polygon. Similarly, in Fig. 2(d), we randomly deploy the Douglas-Peucker algorithm[58] to find the circumscribed or inscribed polygons that fit the object boundaries.

3.3 Dataset statistics

This section discusses several vital statistics of our three SUN-SEG sub-datasets for better illustration. For more details on the data split of SUN-SEG, refer to Section 4.4.1.

Center bias. Unlike general object detection, medical images usually share a higher center bias since the targets are often not in the center of an image. To depict the degree of center bias[59], we compute the average distribution of each dataset′s overall ground-truth map. Figs. 3, 4(a) and 4(b) show that the three sub-datasets of SUN-SEG have a lower center bias than CVC-300 and CVC-612 datasets.

Polyp size. Colonoscopy is an ego-motion situation instead of shooting moving targets (i.e., stuff and things) through fixed cameras in the general domain. As a result, the scale variation of polyps and the irregular movement of the camera causes the different sizes of polyps. The polyps partly or even fully disappear in the view. Fig. 4(c) shows the comparison of polyp scales at five different VPS datasets.

Global/Local contrast. To demonstrate how difficult a colon polyp is to identify, in Fig. 4(d), we describe it quantitatively using the global and local contrast strategy[60].

4 VPS baseline

This section first clarifies the formulation of the VPS

Table 2 List of ten types of visual attributes (ATTR.) and their descriptions

ATTR.	DESCRIPTION
SI	Surgical instruments. The endoscopic surgical procedures involve the positioning of instruments, such as snares, forceps, knives, and electrodes.
IB	Indefinable boundaries. The foreground and background areas around the object have a similar color.
HO	Heterogeneous object. Object regions have distinct colors.
GH	Ghosting. The object has an anomalous RGB-colored boundary due to a fast-moving or insufficient refresh rate.
FM	Fast-motion. The average per-frame object motion in a clip, computed as the Euclidean distance of polyp centroids between consecutive frames, is larger than 20 pixels.
SO	Small object. The average ratio between the object size and the image area in a clip is smaller than 0.05.
LO	Large object. The average ratio between the object size and the image area in a clip is larger than 0.15.
OC	Occlusion. The polyp object becomes partially or fully occluded.
OV	Out-of-view. The polyp object is partially clipped by the image boundaries.
SV	Scale-variation. The average area ratio among any pair of bounding boxes enclosing the target object in a clip is smaller than 0.5.

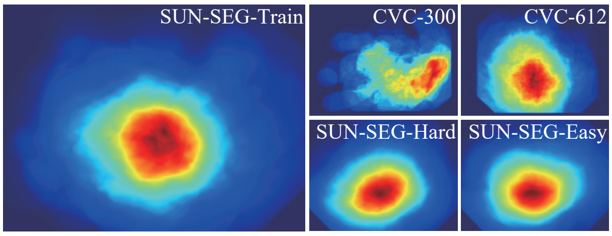
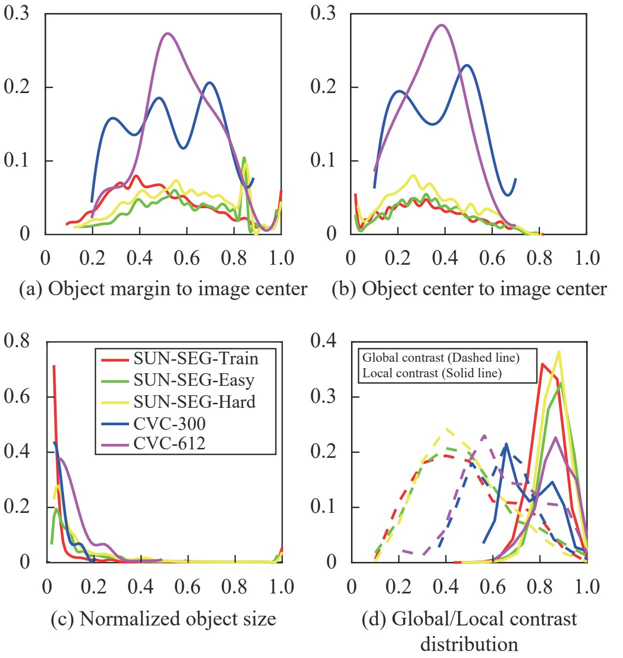


Fig. 3 Calculation of center bias[59] on CVC-300, CVC-612, and our SUN-SEG-Train/SUN-SEG-Easy/SUN-SEG-Hard



(a) Object margin to image center (b) Object center to image center

(c) Normalized object size (d) Global/Local contrast distribution

Fig. 4 Statistic curves among existing VPS datasets (CVC-300 & CVC-612) and our SUN-SEG-Train/SUN-SEG-Easy/SUN-SEG-Hard. Note that the horizontal and vertical axes denote the frequency and their statistic values, respectively. These curves present the diversity of our dataset.

task in Section 4.1. Then, we describe the details of PNS+, including the normalized self-attention block (Section 4.2), global-to-local learning strategy (Section 4.3), and implementation details (Section 4.4).

4.1 Task formulation

We mainly focus on the task of video polyp segmentation, which could be defined as a binary-class video object segmentation task, i.e., identifying polyp and non-polyp areas. Specifically, our goal is to render a model to assign a probability prediction (i.e., a non-binary mask ranging from 0 to 1) for every pixel of each frame. Besides, we leave other types of tasks for future exploration, such as video polyp detection.

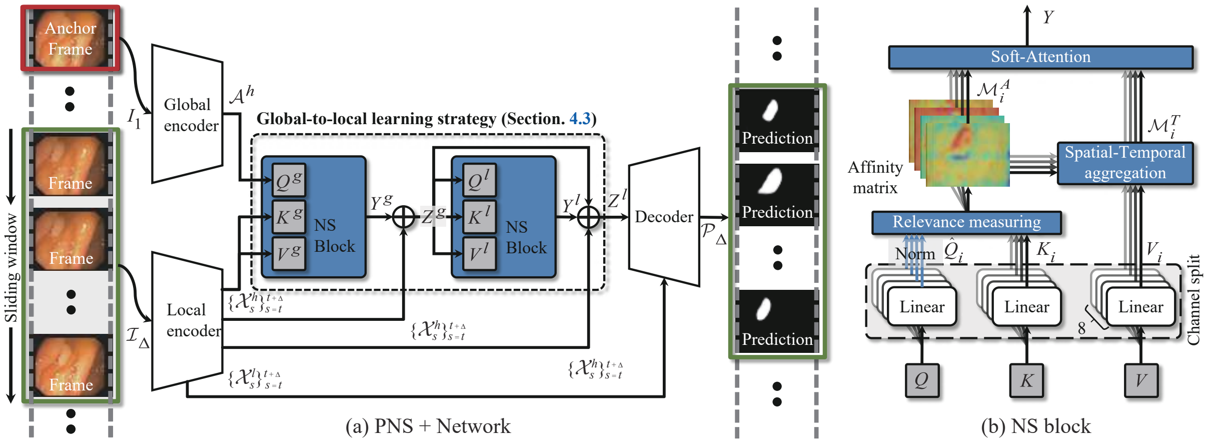
4.2 Normalized self-attention block

Recently, the self-attention mechanism[61] has been widely exploited in many popular computer vision tasks. Our initial studies found that introducing the original self-attention mechanism to the VPS task does not achieve satisfactory results (high accuracy and speed) due to the multiscale property of polyps that are captured at various shooting angles and speeds. Directly utilizing the naive self-attention scheme, such as the non-local network[61], incurs a high computational cost, limiting the inference speed. As shown in Fig. 5 (right), we propose a normalized self-attention (NS) block, which is motivated by the fact that dynamically updating the receptive field is important for self-attention-based networks. The NS block involves five key steps, which are detailed as follows.

4.2.1 Enhanced rules

Motivated by the recent video salient object detection model[62], we utilize three strategies, i.e., channel split rule, query-dependent rule, and normalization rule, to reduce the computational cost and improve the accuracy.

Channel split rule. Specifically, given three candidate features (i.e., query feature Q, key feature K, and value feature V) with the size of $\mathbf{R}^{T\times H\times W\times C}$, we utilize three linear embedding functions $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ to generate the corresponding attention features. These functions can be implemented by a convolutional layer



(a) PNS + Network (b) NS block

Fig. 5 Pipeline of the proposed (a) PNS+ network, which is based on (b) the normalized self-attention (NS) block.

with a kernel size of 1×1×1[61]. Note that T, H, W, and C denote the number of frames, height, width, and channels of the given feature, respectively. This rule can be expressed as

$$Q_i = \mathcal{F}^G\langle\theta(Q)\rangle,\ \ K_i = \mathcal{F}^G\langle\phi(K)\rangle,\ \ V_i = \mathcal{F}^G\langle g(V)\rangle \quad (1)$$

where the function $\mathcal{F}^G$ denotes the operation in which we split each attention feature into N groups along the channel dimension, resulting in three disparate features: query Q_i, key K_i, and value V_i, where $i=\{1,2,\cdots,N\}$. Thus, the shape of the above three split features is $\mathbf{R}^{T\times H\times W\times \frac{C}{N}}$.

Query-dependent rule. To model the spatial-temporal relationship among consecutive frames, we need to measure the similarity between the split query features $\{Q_i\}_{i=1}^N$ and split key features $\{K_i\}_{i=1}^N$. Inspired by [62], we introduce N relevance measuring (i.e., query-dependent rule) blocks to compute the spatial-temporal affinity matrix for the constrained neighborhood of the target pixel. Rather than computing the response between a query position and a key feature at all positions, as done in [61], the relevance measuring block can capture more relevance regarding the target object within T frames. More specifically, we get the corresponding constrained neighborhood in K_i for the query pixel $\boldsymbol{X}^q$ of Q_i in position (x,y,z), which can be obtained by a point sampling function $\mathcal{F}^S$. This is formulated as

$$\mathcal{F}^S\langle\boldsymbol{X}^q, K_i\rangle = \sum_{m=x-kd_i}^{x+kd_i}\ \sum_{n=y-kd_i}^{y+kd_i}\ \sum_{t=1}^{T} K_i(m,n,t) \quad (2)$$

where $1\le x\le H$, $1\le y\le W$, $1\le z\le T$ and $\mathcal{F}^S\langle\boldsymbol{X}^q, K_i\rangle\in\mathbf{R}^{T(2k+1)^2\times\frac{C}{N}}$. Thus, the size of the constrained neighborhood will depend on the various spatial-temporal receptive fields with different kernel sizes k, dilation rate d_i at the i-th group, and frame number T, respectively.

Normalization rule. However, the internal covariate shift problem[63] exists in the feed-forward of input Q_i, causing the layer parameters to not dynamically adapt to the next mini-batch. Thus, we maintain a fixed distribution for Q_i via

$$\hat{Q}_i = \mathrm{Norm}(Q_i) \quad (3)$$

where $\mathrm{Norm}(\cdot)$ is implemented by the layer normalization[64] along the temporal dimension.

4.2.2 Relevance measuring

The affinity matrix $\mathcal{M}_i^A$ measures the similarity of target pixels and their surrounding spatial-temporal contents in an adaptive point sampling manner (refer to (2)). It is defined as

$$\mathcal{M}_i^A = \mathrm{Softmax}\left(\frac{\hat{Q}_i\mathcal{F}^S\langle\hat{\boldsymbol{X}}^q, K_i\rangle^{\mathrm{T}}}{\sqrt{C/N}}\right),\ \text{when}\ \hat{\boldsymbol{X}}^q\in\hat{Q}_i \quad (4)$$

where $\mathcal{M}_i^A\in\mathbf{R}^{THW\times T(2k+1)^2}$. $\sqrt{C/N}$ is a scaling factor to balance the multi-head attention.

4.2.3 Spatial-temporal aggregation

Similar to relevance measuring, we also compute the spatial-temporally aggregated features $\mathcal{M}_i^T\in\mathbf{R}^{THW\times\frac{C}{N}}$ within the constrained neighborhood during temporal aggregation. It is calculated by

$$\mathcal{M}_i^T = \mathcal{M}_i^A\mathcal{F}^S\langle\boldsymbol{X}^a, V_i\rangle,\ \text{when}\ \boldsymbol{X}^a\in\mathcal{M}_i^A. \quad (5)$$

4.2.4 Soft-attention

We utilize a soft-attention block to synthesize features from the group of affinity matrices $\mathcal{M}_i^A$ and aggregated features $\mathcal{M}_i^T$. During the synthesis process, relevant spatial-temporal patterns should be enhanced, while less relevant ones should be suppressed. We first concatenate a group of affinity matrices $\mathcal{M}_i^A$ along the channel dimension to generate $\mathcal{M}^A$. The soft-attention map $\mathcal{M}^S$ is computed by

$$\mathcal{M}^S\in\mathbf{R}^{THW\times 1} = \mathrm{Max}(\mathcal{M}^A) \quad (6)$$

where $\mathcal{M}^A\in\mathbf{R}^{THW\times T(2k+1)^2N}$ and the $\mathrm{Max}(\cdot)$ function computes the channel-wise maximum value. We then concatenate a group of the spatial-temporally aggregated features $\mathcal{M}_i^T$ along the channel dimension to generate $\mathcal{M}^T$.

4.2.5 Normalized self-attention

Finally, our normalized self-attention block, i.e., the function $\mathrm{NS}(\cdot,\cdot,\cdot)$, is defined as

$$Y\in\mathbf{R}^{T\times H\times W\times C} = \mathrm{NS}(Q,K,V) = (\mathcal{M}^T\boldsymbol{W}_T)\circledast\mathcal{M}^S \quad (7)$$

where $\boldsymbol{W}_T$ is the learnable weight and $\circledast$ denotes the channel-wise Hadamard product.

4.3 Global-to-local learning

Observation. By establishing the non-local connections for the given features, the proposed NS block, as in Section 4.2, shows the promising potential for learning short-term spatial-temporal dependencies. However, this mechanism still struggles to model long-term spatial-temporal dependencies due to limited computational resources, i.e., the network can only process a piece of frames within a limited time.

In contrast to our conference version, PNSNet[4], we propose a novel global-to-local learning paradigm, which realizes both long-term and short-term spatial-temporal propagation at an arbitrary temporal distance, yielding a simple but efficient framework, PNS+. Specifically, it appends a spatial-temporal learning pathway at a global temporal level, naturally introducing long-term dependencies into the network. We describe this strategy via the following five steps: a global encoder (Section 4.3.1), a local encoder (Section 4.3.2), the global spatial-temporal modeling (Section 4.3.3), the global-to-local propagation (Section 4.3.4), and the decoder/objectiveness (Section 4.3.5).

4.3.1 Global encoder

Our strategy employs the first frame $I_1 \in \mathbf{R}^{H' \times W' \times 3}$ as an anchor (i.e., global reference). The dependency will be calculated between the anchor frame and the sampled consecutive frames within a sliding window. Following PraNet[48], we use the same backbone, Res2Net-50[65], to extract the feature in the conv4_6 layer. To alleviate the computational burden, we adopt an RFB-like[66] module to reduce the channel dimension of the extracted feature and generate the anchor feature $\mathcal{A}^h \in \mathbf{R}^{H^h \times W^h \times C^h}$.

4.3.2 Local encoder

The local encoder takes a piece of consecutive frames $\mathcal{I}_\Delta = \{I_s\}_{s=t}^{t+\Delta} \in \mathbf{R}^{H' \times W' \times 3}$ $(t > 1)$ from a sliding window as input. Similar to the global encoder, we leverage the Res2Net-50 backbone to extract two groups of short-term features from the conv3_4 and conv4_6 layers and use channel reduction to generate the low-level $\{\mathcal{X}_s^l\}_{s=t}^{t+\Delta} \in \mathbf{R}^{H^l \times W^l \times C^l}$ and high-level $\{\mathcal{X}_s^h\}_{s=t}^{t+\Delta} \in \mathbf{R}^{H^h \times W^h \times C^h}$ short-term features. We set $H^l = H'/4$, $W^l = W'/4$, $C^l = 24$, $H^h = H'/8$, $W^h = W'/8$, and $C^h = 32$ as the default implementation.

4.3.3 Global spatial-temporal modeling

As shown in Fig. 5, we leverage the first NS block to model the long-term relationship at an arbitrary temporal distance, which requires a four-dimensional temporal feature as input; therefore, we have

$$\begin{aligned} &\tilde{\mathcal{X}}^h \in \mathbf{R}^{\Delta \times H^h \times W^h \times C^h} \Leftarrow \{\mathcal{X}_s^h\}_{s=t}^{t+\Delta} \in \mathbf{R}^{H^h \times W^h \times C^h} \\ &\tilde{\mathcal{A}}^h \in \mathbf{R}^{1 \times H^h \times W^h \times C^h} \Leftarrow \mathcal{A}^h \in \mathbf{R}^{H^h \times W^h \times C^h} \end{aligned} \qquad (8)$$

where $\Leftarrow$ denotes reshaping the candidate features into the temporal form to yield a four-dimensional tensor. Then, as for the first NS block formulated in (7), we employ the anchor feature as a query entry (i.e., $Q^g = \tilde{\mathcal{A}}^h$) and the high-level short-term feature as the key and value entries (i.e., $K^g = \tilde{\mathcal{X}}^h$ & $V^g = \tilde{\mathcal{X}}^h$). Intuitively, we aim to build pixel-wise similarities between the anchor and high-level short-term features, which could be viewed as the modeling of global spatial-temporal dependencies. It is defined as

$$Z^g \in \mathbf{R}^{\Delta \times H^h \times W^h \times C^h} = \mathrm{NS}(\tilde{\mathcal{A}}^h, \tilde{\mathcal{X}}^h, \tilde{\mathcal{X}}^h) \oplus \tilde{\mathcal{X}}^h \qquad (9)$$

where $\oplus$ denotes the element-wise addition of residual operation[67]. This operation provides better convergence stability of interior gradient propagation within the first NS block, allowing it to easily be plugged into the pre-trained networks.

4.3.4 Global-to-local propagation

Furthermore, we desire to propagate the long-term dependency Z^g into a local neighborhood (i.e., frames in a sliding window). Thus, we serve Z^g as the input entries of the second NS block as in (7), i.e., query $Q^l = Z^g$, key $K^l = Z^g$, and value $V^l = Z^g$. We have

$$Z^l = \mathrm{NS}(Z^g, Z^g, Z^g) \oplus Z^g \oplus \tilde{\mathcal{X}}^h. \qquad (10)$$

In this way, the introduced two residual connections can maintain the interior gradient stability (i.e., $\oplus Z^g$) and exterior gradient stability (i.e., $\oplus \tilde{\mathcal{X}}^h$) of the second NS block.

4.3.5 Decoder and objectiveness

Finally, we combine the low-level short-term feature $\mathcal{X}_s^l$ from the local encoder and the spatial-temporal feature Z^l from the second NS block with a two-stage UNet-alike decoder $\mathcal{F}^D$. Before the combination, we recover the feature Z^l back to the spatial form, i.e., $\{Z_s^l\}_{s=t}^{t+\Delta}$. The prediction from the decoder is computed with

$$\mathcal{P}_\Delta = \{P_s\}_{s=t}^{t+\Delta} = \mathcal{F}^D \langle \{\mathcal{X}_s^l\}_{s=t}^{t+\Delta}, \{Z_s^l\}_{s=t}^{t+\Delta} \rangle. \qquad (11)$$

To this end, given a prediction P_s and the corresponding ground-truth (GT) G_s at timestamp s, we utilize a binary cross-entropy loss for optimization, which is formulated as

$$\mathcal{L}_{bce} = -\sum [G_s \log(P_s) + (1 - G_s)\log(1 - P_s)]. \qquad (12)$$

4.4 Implementation details

4.4.1 Datasets

We split 40% of the SUN-SEG data for training, i.e., SUN-SEG-Train with 112 clips (19 544 frames). The rest of the data are all used for testing, including SUN-SEG-Easy with 119 clips (17 070 frames) and SUN-SEG-Hard with 54 clips (12 522 frames) according to difficulty levels in each pathological category. Specifically, two colonoscopy scenarios (i.e., seen and unseen[3]) are included in the above two testing datasets: SUN-SEG-Easy (seen: 33 clips & unseen: 86 clips) and SUN-SEG-Hard (seen: 17 clips & unseen: 37 clips) for more fine-grained experimental analyses.

4.4.2 Training details

We train our model using the SUN-SEG-Train dataset on the server platform equipped with an Intel Xeon (R) CPU E5-2690v4×24 and four NVIDIA Tesla V100 GPUs with 16 GB of memory each. The ImageNet pre-trained weights of Res2Net-50[65] are loaded before training, and other newly-added layers are with Kaiming initialization. We set the batch size to 24, which takes about 5 hours to reach convergence after 15 epochs. For each mini-batch of data, we select the first frame of a video clip as an anchor, randomly sample five consecutive frames ($\Delta = 5$) from the same clip, and resize them to 256×448. The Adam optimizer's initial learning rate

[3] Seen denotes that the samples in the testing dataset are from the same case in the training set, whereas the unseen indicates that the scenario do not exist in the training set.

and weight decay are set to 3×10^{-4} and 1×10^{-4}, respectively. We set the number of attention groups to $N=4$ as default. For the first NS block, we set the kernel size $k=3$ and the dilation rate $d_i=\{3,4,3,4\}$ to capture more long-term representations with a larger receptive field. For the second one, we set the kernel size $k=3$ and reduce the dilation rate $d_i=\{1,2,1,2\}$ to mainly focus on short-term relationships.

4.4.3 Inference stage

We evaluate PNS+ on SUN-SEG-Easy and SUN-SEG-Hard with both seen and unseen scenarios. Similar to the training phase, during inference, we select the first frame as an anchor, sample five video frames ($\Delta=5$) from a video clip, and resize them to 256×448. For the final prediction, we use the network's output $\mathcal{P}_\Delta$ followed by a Sigmoid function. The proposed PNS+ achieves a super real-time inference speed of 170 fps on a single V100 GPU without any heuristic post-processing techniques, such as DenseCRF[68].

5 VPS benchmark

5.1 Evaluation protocols

5.1.1 Competitors

We elaborately select eight typical video-based object/polyp segmentation methods, including COSNet[69], MAT[70], PCSA[62], 2/3D[2], AMD[71], DCF[72], FSNet[73], and PNSNet[4]. We also add five image-based object/polyp segmentation methods to validate the effectiveness of per-frame prediction ability, including UNet[34], UNet++[35], ACSNet[38], PraNet[48], and SANet[41]. For a fair comparison, all the competitors utilize the same training dataset as our PNS+ and reach convergence under their default settings. It is worth noting that this paper focuses only on the positive cases (with poly) in our SUN-SEG dataset and leaves negative cases (without polyp) for future work.

5.1.2 Evaluation metrics

To provide a deeper insight into the model performance, we use the following six different metrics for model evaluation between prediction P_s and ground-truth G_s at timestamp s, including: 1) The Dice coefficient ($\text{Dice}=2\times|P_s\cap G_s|/|P_s\cup G_s|$) measures the similarity between prediction and ground-truth mask and penalizes for the false-positive/false-negative predictions. The operators $\cap$, $\cup$, and $|\cdot|$ denote the intersection, union, and the number of pixels in an area, respectively. 2) The pixel-wise sensitivity ($\text{Sen}=|P_s\cap G_s|/|G_s|$) is used to evaluate the true positive prediction of overall lesion areas. Since the goal of a colonoscopy is to screen the polyps with a low polyp missing rate, people who have the polyps should be highly likely to be identified. As a result, penalizing the false-negative prediction can be done by adopting sensitivity, which refers to the method's ability to correctly detect polyps. 3) Being the harmonic mean of precision and recall that is weighted by β, F-measure[74] ($F_\beta=(1+\beta^2)\times\text{Prc}\times\text{Rcl}/(\beta^2\times(\text{Prc}+\text{Rcl}))$) is widely used in measuring binary masks by combining precision ($\text{Prc}=|P_s\cap G_s|/|P_s|$) and recall ($\text{Rcl}=|P_s\cap G_s|/|G_s|$) for more comprehensive evaluation. 4) As suggested by [75, 76], the weighted F-measure[77]($F_\beta^w=(1+\beta^2)\times\text{Prc}^w\times\text{Rcl}^w/(\beta^2\times(\text{Prc}^w+)\text{Rcl}^w)$): amend the "Equal-importance flaw" in Dice and F_β, providing more reliable evaluation results. Following [78], we set the factor β^2 of F_β and F_β^w as 0.3 and 1, respectively. 5) Different from the above pixel-wise metrics, structure measure[79] ($\mathcal{S}_\alpha=\alpha\times\mathcal{S}_o(P_s,G_s)+(1-\alpha)\times\mathcal{S}_r(P_s,G_s)$) is used to measure the structural similarity at object-aware $\mathcal{S}_o$ and region-aware $\mathcal{S}_r$, respectively. We use the factor $\alpha=0.5$ as a default. 6) Fan et al.[80] proposed a human visual perception-based metric, enhanced-alignment measure: $E_\phi=(1/(W\times H))\sum_x^W\sum_y^H\phi(P_s(x,y),G_s(x,y))$, where ϕ is the enhanced-alignment matrix, W and H are the width and height of the ground-truth G_s. This metric is inherently suitable for assessing polyps' heterogeneous location and shape in colonoscopy.

As mentioned in Section 4.1, the models generate continuous floating predictions; thus, we need to threshold the floating value into binary ones ranging from 0 to 255. Specifically, we provide the maximum value of Dice and the mean value of E_ϕ, F_β, and Sen under different thresholds for the binary metrics. Furthermore, we obtain the video-level score by averaging the evaluated results per image at a video clip. Then, we take the average video-level scores as the performance on the whole dataset. The one-key evaluation toolbox is available at https://github.com/GewelsJI/VPS/tree/main/eval.

5.2 Quantitative comparison

Based on the protocols mentioned in Section 5.1, we conduct a comprehensive VPS benchmark on two testing sub-datasets (i.e., SUN-SEG-Easy and SUN-SEG-Hard), which include the following three aspects.

5.2.1 Learning ability

Notably, the image-based models are trained and inferred frame-by-frame. To better unveil the spatial-temporal learning ability on the colonoscopy videos, we conduct two groups of experiments to validate the video-based competitors' ability on two seen sub-datasets. For these sub-datasets shown in Table 3, our PNS+ also outperforms top-1 video-based approaches, e.g., Dice score on SUN-SEG-Easy (Seen): PNSNet (0.861) versus PNS+ (0.888) and F_ϕ^{mn} score on SUN-SEG-Hard (Seen): PNSNet (0.892) versus PNS+ (0.929). The above results suggest that PNS+ has a strong learning ability to segment polyps accurately.

5.2.2 Generalization capability

To validate the model's generalizability, we conduct the experiments on two testing sub-datasets with unseen colonoscopy scenarios. As shown in Table 4, we present the performance comparison with the other latest image- and video-based competitors in six metrics. It shows that

Table 3 Quantitative comparison of two testing sub-datasets with seen colonoscopy scenarios

Model	SUN-SEG-Easy (Seen)				SUN-SEG-Hard (Seen)			
	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice
COSNet[69]	0.845	0.836	0.727	0.804	0.785	0.772	0.626	0.725
MAT[70]	0.879	0.861	0.731	0.833	0.840	0.821	0.652	0.776
PCSA[62]	0.852	0.835	0.681	0.779	0.772	0.759	0.566	0.679
2/3D[2]	0.895	0.909	0.819	0.856	0.849	0.868	0.753	0.809
AMD[71]	0.471	0.526	0.114	0.245	0.480	0.536	0.115	0.231
DCF[72]	0.572	0.591	0.357	0.398	0.603	0.602	0.385	0.443
FSNet[73]	0.890	0.895	0.818	0.873	0.848	0.859	0.755	0.828
PNSNet[4]	0.906	0.910	0.836	0.861	0.870	0.892	0.787	0.823
PNS+	**0.917**	**0.924**	**0.848**	**0.888**	**0.887**	**0.929**	**0.806**	**0.855**

Table 4 Quantitative comparison of two testing sub-datasets with unseen colonoscopy scenarios. "R/T" means to retrain the private model using the code provided by the author. The best values are highlighted in **bold**.

	Model	Publish	Code	SUN-SEG-Easy (Unseen)						SUN-SEG-Hard (Unseen)					
				$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	F_β^{mn}	Dice	Sen
Image-level methods	UNet[34]	MICCAI_{15}	Link	0.669	0.677	0.459	0.528	0.530	0.420	0.670	0.679	0.457	0.527	0.542	0.429
	UNet++[35]	TMI_{18}	Link	0.684	0.687	0.491	0.553	0.559	0.457	0.685	0.697	0.480	0.544	0.554	0.467
	ACSNet[38]	MICCAI_{20}	Link	0.782	0.779	0.642	0.688	0.713	0.601	0.783	0.787	0.636	0.684	0.708	0.618
	PraNet[48]	MICCAI_{20}	Link	0.733	0.753	0.572	0.632	0.621	0.524	0.717	0.735	0.544	0.607	0.598	0.512
	SANet[41]	MICCAI_{21}	Link	0.720	0.745	0.566	0.634	0.649	0.521	0.706	0.743	0.526	0.580	0.598	0.505
Video-level methods	COSNet[69]	TPAMI_{19}	Link	0.654	0.600	0.431	0.496	0.596	0.359	0.670	0.627	0.443	0.506	0.606	0.380
	MAT[70]	TIP_{20}	Link	0.770	0.737	0.575	0.641	0.710	0.542	0.785	0.755	0.578	0.645	0.712	0.579
	PCSA[62]	AAAI_{20}	Link	0.680	0.660	0.451	0.519	0.592	0.398	0.682	0.660	0.442	0.510	0.584	0.415
	2/3D[2]	MICCAI_{20}	R/T	0.786	0.777	0.652	0.708	0.722	0.603	0.786	0.775	0.634	0.688	0.706	0.607
	AMD[71]	NeurIPS_{21}	Link	0.474	0.533	0.133	0.146	0.266	0.222	0.472	0.527	0.128	0.141	0.252	0.213
	DCF[72]	ICCV_{21}	Link	0.523	0.514	0.270	0.312	0.325	0.340	0.514	0.522	0.263	0.303	0.317	0.364
	FSNet[73]	ICCV_{21}	Link	0.725	0.695	0.551	0.630	0.702	0.493	0.724	0.694	0.541	0.611	0.699	0.491
	PNSNet[4]	MICCAI_{21}	Link	0.767	0.744	0.616	0.664	0.676	0.574	0.767	0.755	0.609	0.656	0.675	0.579
	PNS+	**OURS_{22}**	Link	**0.806**	**0.798**	**0.676**	**0.730**	**0.756**	**0.630**	**0.797**	**0.793**	**0.653**	**0.709**	**0.737**	**0.623**

our PNS+ achieves significant improvements by a large margin in comparison with top image-based and video-based approaches, e.g., Dice score on SUN-SEG-Easy (Unseen): ACSNet (0.713) versus 2/3D (0.722) versus PNS+ (0.756) and F_β^w score on SUN-SEG-Hard (Unseen): ACSNet (0.636) versus 2/3D (0.634) versus PNS+ (0.653). Interestingly, we observe that PNSNet drops dramatically on two unseen datasets, which is a sideshow of better generalizability attributed to our newly-proposed global-to-local learning strategy, especially on a clip with a larger time span.

5.2.3 Attribute-based performance

Finally, we analyze the visual attribute-based comparison presented in Table 2. In terms of $\mathcal{S}_\alpha$ score, Table 5 shows that our PNS+ consistently outperforms other rivals in four attributes (i.e., IB, GH, FM, and SV). More specifically, as shown in Table 5, most methods cannot address the VPS tasks with the IB attribute since the colon polyps always have fuzzy boundaries. In contrast, PNS+ achieves the best score ($\mathcal{S}_\alpha = 0.667$) on this challenging IB attribute of SUN-SEG-Easy (Unseen). This discovery is also consistent with the results shown in Fig. 6. Similarly, the SO attribute also presents lower scores (e.g., SUN-SEG-Easy (Unseen): $\mathcal{S}_\alpha = 0.667$), which indicates that these two attributes are the most challenging issues in colonoscopy. On the contrary, the HO and LO attributes consistently sustain higher scores than other attributes, making polyps easier to detect. This phenomenon meets our expectations since there is less distribution bias for these relatively easy scenarios. We refer the reader to Section 5.5 for a more visualized analysis of challenging cases.

Table 5 Visual attributes-based performance on SUN-SEG-Easy/SUN-SEG-Hard (Unseen) in terms of structure measure ($\mathcal{S}_\alpha$) score

	SUN-SEG-Easy (Unseen)										SUN-SEG-Hard (Unseen)									
	SI	IB	HO	GH	FM	SO	LO	OC	OV	SV	SI	IB	HO	GH	FM	SO	LO	OC	OV	SV
UNet[34]	0.675	0.548	0.768	0.715	0.633	0.593	0.648	0.670	0.643	0.620	0.618	0.619	0.663	0.676	0.713	0.689	0.633	0.658	0.659	0.658
UNet++[35]	0.701	0.542	0.782	0.739	0.647	0.591	0.678	0.683	0.665	0.617	0.654	0.604	0.665	0.696	0.714	0.681	0.660	0.676	0.677	0.678
ACSNet[38]	0.789	0.612	0.896	0.820	0.704	0.663	0.787	0.770	0.759	0.705	0.770	0.681	0.828	0.795	0.817	0.738	0.810	**0.828**	0.806	0.759
PraNet[48]	0.745	0.585	0.821	0.772	0.673	0.611	0.722	0.722	0.703	0.653	0.673	0.635	0.725	0.720	0.755	0.691	0.666	0.714	0.708	0.703
SANet[41]	0.724	0.582	0.854	0.760	0.676	0.615	0.703	0.701	0.711	0.680	0.658	0.565	0.738	0.709	0.760	0.692	0.733	0.729	0.727	0.693
COSNet[69]	0.663	0.531	0.786	0.684	0.610	0.549	0.637	0.648	0.613	0.617	0.641	0.593	0.727	0.668	0.690	0.637	0.694	0.707	0.666	0.625
MAT[70]	0.772	0.664	0.873	0.789	0.706	**0.691**	0.755	0.738	0.746	0.715	**0.772**	0.701	0.801	0.776	0.782	0.780	0.791	0.795	0.789	0.750
PCSA[62]	0.676	0.563	0.759	0.708	0.628	0.610	0.634	0.662	0.656	0.616	0.656	0.591	0.692	0.683	0.706	0.671	0.612	0.677	0.665	0.663
2/3D[2]	0.809	0.625	**0.899**	0.835	0.728	0.667	**0.820**	**0.783**	0.778	0.719	0.768	0.662	**0.865**	0.784	0.797	0.737	**0.853**	0.827	**0.808**	0.765
AMD[71]	0.476	0.461	0.471	0.481	0.484	0.466	0.447	0.467	0.442	0.498	0.471	0.468	0.447	0.473	0.468	0.469	0.453	0.487	0.462	0.481
DCF[72]	0.465	0.485	0.479	0.505	0.541	0.495	0.362	0.484	0.492	0.495	0.441	0.508	0.422	0.498	0.587	0.556	0.351	0.470	0.494	0.540
FSNet[73]	0.719	0.603	0.810	0.752	0.694	0.632	0.686	0.711	0.691	0.665	0.662	0.648	0.743	0.713	0.774	0.723	0.701	0.728	0.728	0.694
PNSNet[4]	0.789	0.592	0.871	0.820	0.723	0.619	0.768	0.749	0.751	0.705	0.746	0.631	0.803	0.780	0.778	0.743	0.805	0.790	0.794	0.758
PNS+	**0.819**	**0.667**	0.883	**0.844**	**0.738**	0.690	0.796	0.782	**0.798**	**0.734**	0.770	**0.703**	0.817	**0.801**	**0.823**	**0.793**	0.792	0.808	0.807	**0.795**

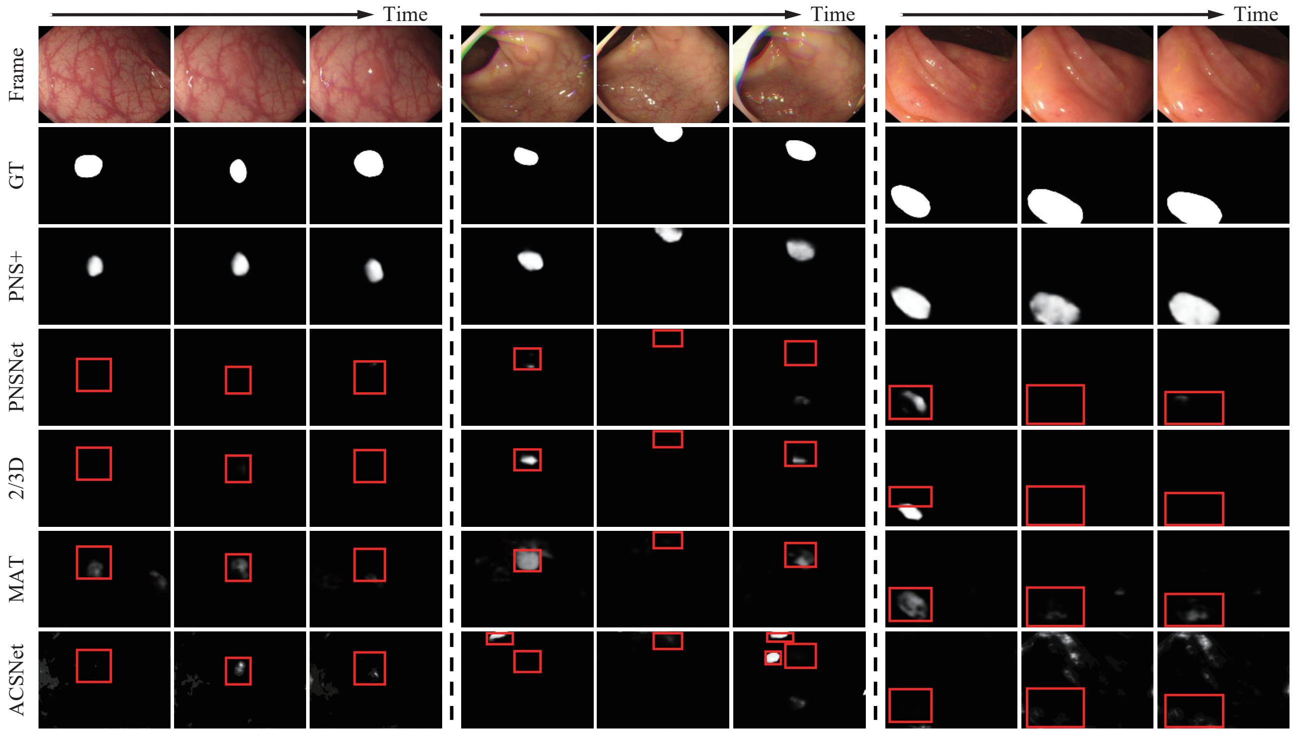


Fig. 6 Qualitative visualization of the proposed PNS+ and four representative competitors on three sequences (from left to right: case14_3, case30, and case3_2). The red boxes indicate the wrong or missing predictions. We refer the reader to the project page for a complete dynamic comparison.

5.3 Qualitative comparison

As shown in Fig. 6, we present visual results on three video clips of four typical models (i.e., PNSNet, 2D/3D, MAT, and ACSNet) and our PNS+. In the last four rows, the competitors fail to generate complete segmentation results for the polyps that share the same camouflaged texture with the background. In contrast, in the 3rd row, our model can accurately locate and segment polyps in a challenging situation, i.e., polyps with different sizes and homogeneous textures.

5.4 Ablation studies

To validate the effectiveness of our core designs, we conduct extensive ablation studies and summarize the results in Table 6.

5.4.1 Contribution of base network

We initialize a UNet-like variant #01 via leveraging the Res2Net-50[65] backbone, which can be viewed as an image-based approach to generate per-frame predictions. We observe that #OUR significantly improves the performance ($\mathcal{S}_\alpha$: +7.7%) on SUN-SEG-Easy (Unseen).

5.4.2 Contribution of channel split

To discover the best setting for the channel split rule as in Equ. (1), we instantiate four variants with four different channel split numbers: #02 ($N=1$), #03 ($N=2$), #04 ($N=4$), and #05 ($N=8$). These results show that small (#02 & #03) and large (#05) channel split numbers may harm the channel-level information by collapsing the knowledge in a different channel. In contrast, we adopt the moderate scale (#04: $N=4$) with the best performance on SUN-SEG-Hard (Unseen) (e.g., Dice: 2.7%↑) when compared to variant #05. Such a trade-off scale would exert our model focusing on the polyp-related attention while suppressing the irrelevant cues.

5.4.3 Contribution of soft-attention

We further ablate soft-attention and observe that #04 with the soft-attention block is generally better than #06 without it on SUN-SEG-Easy (Unseen): 1.9%↑ in terms of the Dice score. Such an improvement suggests that introducing the soft-attention operation to synthesize the relationship between aggregation feature and affinity matrix is necessary for increasing performance.

5.4.4 Effectiveness of normalization

We also study the improvement of the normalization operation by comparing #04 with #07. We observe that #04 generally outperforms #07 on SUN-SEG-Hard (Unseen) (e.g., Dice: 4.1%↑). It shows that the layer normalization along the temporal dimension could alleviate the internal covariate shift problem by fixing the distribution of query entries in the attention mechanism.

5.4.5 Different learning strategies

Finally, we examine the effectiveness of the proposed learning strategy, as proposed in Section 4.3, by deriving three variants, including #08 (L→L: local-to-local), #09 (L→G: local-to-global), #10 (G→G: global-to-global), and #Our (G→L: global-to-local). For example, variant #09 combines local spatial-temporal cues and introduces global ones, termed a local-to-global (L→G) strategy. #08 will dramatically decrease on SUN-SEG-Easy (Unseen) ($\mathcal{S}_\alpha$: 5.8%↓) when focusing on the local cues due to a lack of global context. On the other hand, if only focusing on the global information, the performance of variant #10 will drop on SUN-SEG-Hard (Unseen), e.g., F_β^w: 5.4%↓. In contrast, #Our with the global-to-local strategy outperforms variant #09 on SUN-SEG-Hard (Unseen), e.g., F_β^w: 3.5%↑, since propagating long-term cues into short-term neighbors.

We further validate the effectiveness of the global-to-local learning strategy via visualizing the key dataflows. As shown in Fig. 7, the first and second columns present the anchor feature $\mathcal{A}^h$ extracted from the global encoder and the spatial-temporal feature Z^l from the second NS block, respectively. Note that the current frame I_s is randomly selected from consecutive frames I_Δ. It shows that our PNS+ can propagate the long-term dependency with the assistance of the anchor frame I_1, though the current frame I_s is hard to recognize due to indefinable boundaries (i.e., IB attribute). Of note, as in the rightmost column of Fig. 6, the PNSNet fails to locate the polyp since it does not use a global-to-local learning strategy. Compared to it, our PNS+ successfully detects the polyp by exploiting the global reference of the anchor frame.

5.5 Issues and challenges

This section discusses some common issues within challenging attributes, whose visualization results are

Table 6 Ablation studies for the core designs of the proposed PNS+. See Section 5.4 for detailed analyses

No.	VARIANTS					SUN-SEG-Easy (Unseen)				SUN-SEG-Hard (Unseen)			
	Base	N	Soft	Norm	Strategy	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ^{mn}	F_β^w	Dice
#01	✓	–	–	–	–	0.729	0.718	0.571	0.616	0.726	0.720	0.559	0.603
#02	✓	1	✓	✓	L	0.782	0.766	0.631	0.722	0.783	0.775	0.629	0.715
#03	✓	2	✓	✓	L	0.773	0.760	0.625	0.720	0.785	0.784	0.631	0.719
#04	✓	4	✓	✓	L	0.786	0.777	0.651	0.741	0.792	0.789	0.649	0.735
#05	✓	8	✓	✓	L	0.774	0.762	0.627	0.724	0.775	0.774	0.619	0.708
#06	✓	4	–	✓	L	0.782	0.775	0.639	0.722	0.785	0.786	0.637	0.715
#07	✓	4	✓	–	L	0.755	0.752	0.587	0.705	0.754	0.751	0.579	0.694
#08	✓	4	✓	✓	L→L	0.748	0.717	0.577	0.705	0.760	0.741	0.587	0.693
#09	✓	4	✓	✓	L→G	0.788	0.780	0.645	0.741	0.776	0.768	0.618	0.715
#10	✓	4	✓	✓	G→G	0.778	0.763	0.627	0.726	0.767	0.753	0.599	0.694
#Our	✓	4	✓	✓	G→L	**0.806**	**0.798**	**0.676**	**0.756**	**0.797**	**0.793**	**0.653**	**0.737**

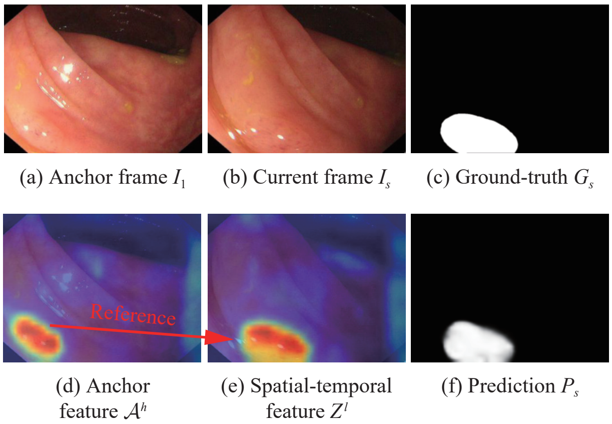


Fig. 7 Feature visualization of key dataflows. The red arrow denotes using the anchor feature $\mathcal{A}^h$ to guide the representation of spatial-temporal frame Z^l. For more details, refer to Section 5.4.5.

presented in Fig. 8. Of note, VPS is a newly-emerging and challenging track in medical imaging, and its overall accuracy is not high enough. We observe that existing cutting-edge models (i.e., ACSNet and 2/3D) and our baseline model (PNS+) still lack sufficient robustness in particular cases in the LO, HO, SI, GH, and SV attributes. As for the HO (3rd column) and LO (8th column) attributes, three models fail to capture the whole polyp due to significant appearance changes. Besides, the false-positive/false-negative predictions (marked with red boxes) on the surgical instrument (1st column) and the optical flares (4th column) indicate that these models could not learn semantics without perceiving the accurate polyp-related representation in such a hard case. Moreover, the misidentifications for the SV attribute (last column) are caused by the insufficient diversity of polyp sizes in the training set. The aforementioned drawbacks inspire us to explore more robust learning paradigms to improve the accuracy of VPS.

We also observe that three models consistently fail to locate lesion regions that share a similar color to the intestinal wall or are too small to be detected. Thus, there is a large room for improving the detection ability in IB and SO attributes via camouflaged pattern discovery techniques[81, 82]. Last but not least, lacking temporal-wise understanding will lead to a false prediction of the FM, OV, and OC attributes. Taking OV and OC, e.g., exploiting temporal cues more thoroughly should mitigate the performance degradation results from the occlusion of the intestinal wall or the image boundary, since the occlusion is not continuous in the entire video clip. In summary, these challenging cases are common difficulties that other methods face and cause severe performance degradation that deserves further exploration.

6 Potential directions

This section highlights several potential trends for promoting colonoscopy research in the deep era.

High-precision diagnosis. As shown in Table 4, we observe that the leading approaches are still unsatisfactory in our SUN-SEG-Hard (e.g., sensitivity score < 0.63). We argue that the high-precision VPS algorithm would steer clinical medicine in boosting auxiliary diagnostic technologies.

Data-insufficient learning. It is promising to explore efficient learning strategies[83, 84] under limited conditions in specific clinical applications, such as weakly-supervised/un-supervised/self-supervised learning and knowledge distillation.

Privacy-preserving AI. Intelligent VPS systems must safeguard data through the entire life cycle from training to production and governance, which fuels fundamental techniques like federal learning.

Trustworthy AI. How AI-guided decisions are made and what determining factors are involved play a crucial role in understanding the insights of deep networks. In other words, the VPS model should be causal, transpar-

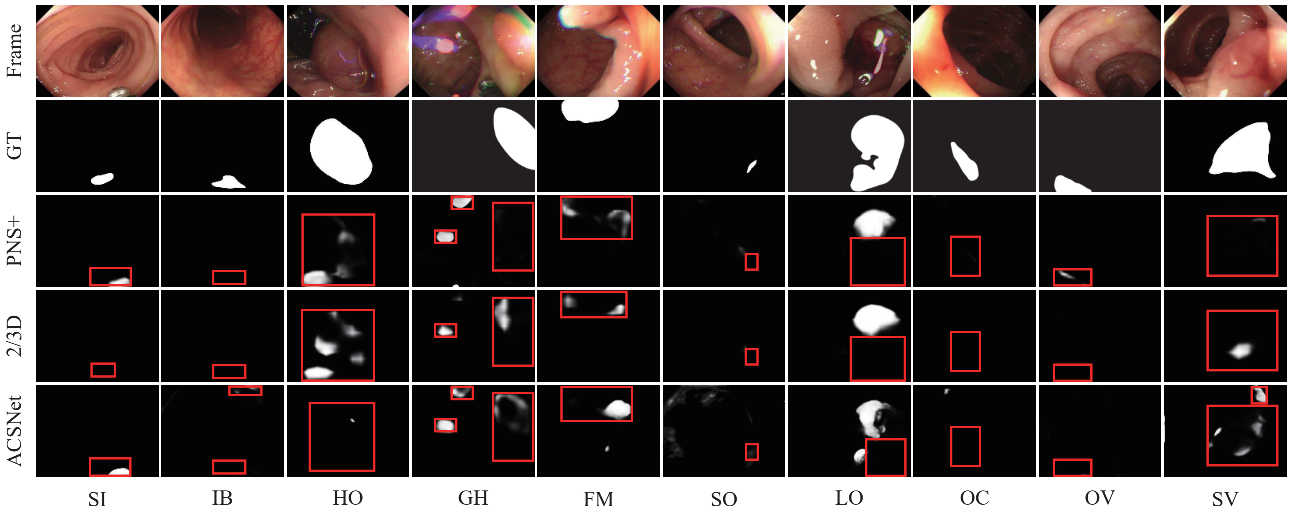


Fig. 8 Challenging samples were taken from ten visual attributes. More analyses can be referred in Section 5.5

ent, explainable, and interactive, which inspires more trusted developments, such as in [85].

The above possible directions listed are still far from being solved for the VPS. Fortunately, several famous works can serve as references, providing it a potential basis to be transferred to the VPS community.

7 Conclusions

This paper presents the first comprehensive study on video polyp segmentation (VPS) from a deep learning perspective. We first introduce a large-scale VPS dataset SUN-SEG via extending the famous SUN-database with diversified annotations, i.e., attribute, object mask, boundary, scribble, and polygon. We then design a simple but efficient baseline, dubbed PNS+, to segment colon polyps from the colonoscopy video. Based on the normalized self-attention block, PNS+ fully exploits long-term and short-term spatial-temporal cues via a novel global-to-local learning strategy. We also contribute the first comprehensive benchmark that contains 13 cutting-edge polyp/object segmentation approaches. Extensive results show that PNS+ achieves the best performance against all these competitors. We conclude by outlining several potential directions for future colonoscopy-related research in the deep learning era. We hope that this work will spur advancements in other closely related medical video analyses.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62072223), and supported by the Natural Science Foundation of Fujian Province, China (No. 2020J01131199). The authors would like to thank the anonymous reviewers and editors for their helpful comments on this manuscript. Besides, we thank Huazhu Fu for his insightful feedback.

Conflicts of Interests

The authors declared that they have no conflicts of interest in this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access

References

[1] J. Bernal, J. Sánchez, F. Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012. DOI: 10.1016/j.patcog.2012.03.002.

[2] J. G. B. Puyal, K. K. Bhatia, P. Brandao, O. F. Ahmad, D. Toth, R. Kader, L. Lovat, P. Mountney, D. Stoyanov. Endoscopic polyp segmentation using a hybrid 2D/3D CNN. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Lima, Peru, pp. 295–305, 2020. DOI: 10.1007/978-3-030-59725-2_29.

[3] M. Misawa, S. E. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, K. Mori. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967, 2021. DOI: 10.1016/j.gie.2020.07.060.

[4] G. P. Ji, Y. C. Chou, D. P. Fan, G. Chen, H. Z. Fu, D. Jha, L. Shao. Progressively normalized self-attention network for video polyp segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 142–152, 2021. DOI: 10.1007/978-3-030-87193-2_14.

[5] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014. DOI: 10.1007/s11548-013-0926-3.

[6] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. 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. DOI: 10.1016/j.compmedimag.2015.02.007.

[7] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, A. Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, vol. 35, no. 9, pp. 2051–2063, 2016. DOI: 10.1109/TMI.2016.2547947.

[8] N. Tajbakhsh, S. R. Gurudu, J. M. Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016. DOI: 10.1109/TMI.

2015.2487997.

[9] Gastrointestinal Image ANAlysis (GIANA) Challenge. [Online], Available: https://endovissub2017-giana.grand-challenge.org/.

[10] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, A. Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, vol. 2017, Article number 4037190, 2017. DOI: 10.1155/2017/4037190.

[11] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz, G. Mavrogenis, A. Nemeth, H. Thorlacius, G. E. Tontini. Kid project: An internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy International Open*, vol. 5, no. 6, pp. E477–E483, 2017. DOI: 10.1055/s-0043-105488.

[12] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, V. P. Plagianakos. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2196–2210, 2018. DOI: 10.1109/TMI.2018.2837002.

[13] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. De Lange, D. Johansen, C. Spampinato, D. T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, P. Halvorsen. KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, Taipei, China, pp. 164–169, 2017. DOI: 10.1145/3083187.3083212.

[14] S. Ali, N. Ghatwary, B. Braden, D. Lamarque, A. Bailey, S. Realdon, R. Cannizzaro, J. Rittscher, C. Daul, J. East. Endoscopy disease detection challenge 2020. [Online], Available: https://arxiv.org/abs/2003.03376, 2020.

[15] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, T. De Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, vol. 7, no. 1, Article number 283, 2020. DOI: 10.1038/s41597-020-00622-y.

[16] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, H. D. Johansen. Kvasir-SEG: A segmented polyp dataset. In *Proceedings of the 26th International Conference on Multimedia Modeling*, Springer, Daejeon, Korea, pp. 451–462, 2020. DOI: 10.1007/978-3-030-37734-2_37.

[17] L. F. Sánchez-Peralta, J. B. Pagador, A. Picón, Á. J. Calderón, F. Polo, N. Andraka, R. Bilbao, B. Glover, C. L. Saratxaga, F. M. Sánchez-Margallo. PICCOLO white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets. *Applied Sciences*, vol. 10, no. 23, Article number 8501, 2020. DOI: 10.3390/app10238501.

[18] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. De Lange, M. A. Riegler, P. Halvorsen. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, vol. 8, no. 1, Article number 142, 2021. DOI: 10.1038/s41597-021-00920-z.

[19] W. Wang, J. G. Tian, C. W. Zhang, Y. H. Luo, X. Wang, J. Li. An improved deep learning approach and its applications on colonic polyp images detection. *BMC Medical Imaging*, vol. 20, no. 1, Article number 83, 2020. DOI: 10.1186/s12880-020-00482-3.

[20] Y. T. Ma, X. J. Chen, K. Cheng, Y. Li, B. Sun. LDPolypVideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 387–396, 2021. DOI: 10.1007/978-3-030-87240-3_37.

[21] K. D. Li, M. I. Fathan, K. Patel, T. X. Zhang, C. C. Zhong, A. Bansal, A. Rastogi, J. S. Wang, G. H. Wang. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *PLoS One*, vol. 16, no. 8, Article number e0255809, 2021. DOI: 10.1371/journal.pone.0255809.

[22] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, K. V. Anonsen, A. Petlund, P. Halvorsen, J. Rittscher, T. De Lange, J. E. East. Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment. [Online], Available: https://arxiv.org/abs/2106.04463.

[23] B. V. Dhandra, R. Hegadi, M. Hangarge, V. S. Malemath. Analysis of abnormality in endoscopic images using combined hsi color space and watershed segmentation. In *Proceedings of the 18th International Conference on Pattern Recognition*, IEEE, Hong Kong, China, pp. 695–698, 2006. DOI: 10.1109/ICPR.2006.268.

[24] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, Y. H. R. Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, 2014. DOI: 10.1109/TMI.2014.2314959.

[25] O. H. Maghsoudi. Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In *Proceedings of the Signal Processing in Medicine and Biology Symposium*, IEEE, Philadelphia, USA, 2017. DOI: 10.1109/SPMB.2017.8257027.

[26] L. Q. Yu, H. Chen, Q. Dou, J. Qin, P. A. 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, 2017. DOI: 10.1109/JBHI.2016.2637004.

[27] W. Tavanapong, J. Oh, M. A. Riegler, M. Khaleel, B. Mittal, P. C. De Groen. Artificial intelligence for colonoscopy: Past, present, and future. *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3950–3965, 2022. DOI: 10.1109/JBHI.2022.3160098.

[28] H. Gammulle, S. Denman, S. Sridharan, C. Fookes. Two-stream deep feature modelling for automated video endo-

scopy data analysis. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Lima, Peru, pp. 742–751, 2020. DOI: 10.1007/978-3-030-59716-0_71.

[29] G. Carneiro, L. Z. C. T. Pu, R. Singh, A. Burt. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis*, vol. 62, Article number 101653, 2020. DOI: 10.1016/j.media.2020.101653.

[30] R. K. Zhang, Y. L. Zheng, C. C. Y. Poon, D. G. Shen, J. Y. W. Lau. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. *Pattern Recognition*, vol. 83, pp. 209–219, 2018. DOI: 10.1016/j.patcog.2018.05.026.

[31] L. Y. Wu, Z. Q. Hu, Y. F. Ji, P. Luo, S. T. Zhang. Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 302–312, 2021. DOI: 10.1007/978-3-030-87240-3_29.

[32] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, D. Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Proceedings of the SPIE 10134, Medical Imaging 2017: Computer-aided Diagnosis*, Orlando, USA, pp. 101–107, 2017. DOI: 10.1117/12.2254361.

[33] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S. M. R. Soroushmehr, N. Karimi, S. Samavi, K. Najarian. Polyp segmentation in colonoscopy images using fully convolutional network. In *Proceedings of the 40th Annual International Conference of the Engineering in Medicine and Biology Society*, Honolulu, USA, pp. 69–72, 2018. DOI: 10.1109/EMBC.2018.8512197.

[34] O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Munich, Germany, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[35] Z. W. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. M. Liang. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020. DOI: 10.1109/TMI.2019.2959609.

[36] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen. ResuNet++: An advanced architecture for medical image segmentation. In *Proceedings of the International Symposium on Multimedia*, IEEE, San Diego, USA, pp. 225–2255, 2019. DOI: 10.1109/ISM46123.2019.00049.

[37] J. F. Zhong, W. Wang, H. S. Wu, Z. K. Wen, J. Qin. PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Lima, Peru, pp. 285–294, 2020. DOI: 10.1007/978-3-030-59725-2_28.

[38] R. F. Zhang, G. B. Li, Z. Li, S. G. Cui, D. H. Qian, Y. Z. Yu. Adaptive context selection for polyp segmentation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Lima, Peru, pp. 253–262, 2020. DOI: 10.1007/978-3-030-59725-2_25.

[39] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, P. Halvorsen. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access*, vol. 9, pp. 40496–40510, 2021. DOI: 10.1109/ACCESS.2021.3063716.

[40] H. S. Wu, J. F. Zhong, W. Wang, Z. K. Wen, J. Qin. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In *Proceedings of AAAI Conference on Artificial Intelligence*, Palo Alto, USA, pp. 2916–2924, 2021.

[41] J. Wei, Y. W. Hu, R. M. Zhang, Z. Li, S. K. Zhou, S. G. Cui. Shallow attention network for polyp segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 699–708, 2021. DOI: 10.1007/978-3-030-87193-2_66.

[42] X. Q. Zhao, L. H. Zhang, H. C. Lu. Automatic polyp segmentation via multi-scale subtraction network. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 120–130, 2021. DOI: 10.1007/978-3-030-87193-2_12.

[43] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, M. Sivaprakasam. PSI-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Berlin, Germany, pp. 7223–7226, 2019. DOI: 10.1109/EMBC.2019.8857339.

[44] R. X. Wang, S. Y. Chen, C. J. Ji, J. P. Fan, Y. Li. Boundary-aware context neural network for medical image segmentation. *Medical Image Analysis*, vol. 78, Article number 102395, 2022. DOI: 10.1016/j.media.2022.102395.

[45] Y. Q. Fang, C. Chen, Y. X. Yuan, K. Y. Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *Proceedings of the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Shenzhen, China, pp. 302–310, 2019. DOI: 10.1007/978-3-030-32239-7_34.

[46] Y. T. Shen, X. Jia, M. Q. H. Meng. HRENet: A hard region enhancement network for polyp segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 559–568, 2021. DOI: 10.1007/978-3-030-87193-2_53.

[47] G. P. Ji, L. Zhu, M. C. Zhuge, K. R. Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, vol. 123, Article number 108414, 2022. DOI: 10.1016/j.patcog.2021.108414.

[48] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Z. Fu, J. B. Shen, L. Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Lima, Peru, pp. 263–273, 2020. DOI: 10.1007/978-3-030-59725-2_26.

[49] T. C. Nguyen, T. P. Nguyen, G. H. Diep, A. H. Tran-Dinh, T. V. Nguyen, M. T. Tran. CCBANet: Cascading context and balancing attention for polyp segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 633–643, 2021. DOI: 10.1007/978-3-030-87193-2_60.

[50] M. J. Cheng, Z. S. Kong, G. L. Song, Y. H. Tian, Y. S. Liang, J. Chen. Learnable oriented-derivative network for polyp segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 720–730, 2021. DOI: 10.1007/978-3-030-87193-2_68.

[51] T. Kim, H. Lee, D. Kim. UACANet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, Chengdu, China, pp. 2167–2175, 2021. DOI: 10.1145/3474085.3475375.

[52] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, H. Z. Fu. Transformers in medical imaging: A survey. [Online], Available: https://arxiv.org/abs/2201.09873, 2022.

[53] Y. D. Zhang, H. Y. Liu, Q. Hu. Transfuse: Fusing transformers and CNNs for medical image segmentation. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention*, Springer, Strasbourg, France, pp. 14–24, 2021. DOI: 10.1007/978-3-030-87193-2_2.

[54] S. H. Li, X. C. Sui, X. D. Luo, X. X. Xu, Y. Liu, R. Goh. Medical image segmentation using squeeze-and-expansion transformers. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Montreal, Canada, pp. 807–815, 2021. DOI: 10.24963/ijcai.2021/112.

[55] W. H. Wang, E. Z. Xie, X. Li, D. P. Fan, K. T. Song, D. Liang, T. Lu, P. Luo, L. Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022. DOI: 10.1007/s41095-022-0274-8.

[56] B. Dong, W. H. Wang, D. P. Fan, J. P. Li, H. Z. Fu, L. Shao. Polyp-PVT: Polyp segmentation with pyramid vision transformers. [Online], Available: https://arxiv.org/abs/2108.06932, 2021.

[57] D. P. Fan, G. P. Ji, G. L. Sun, M. M. Cheng, J. B. Shen, L. Shao. Camouflaged object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 2777–2787, 2020. DOI: 10.1109/CVPR42600.2020.00285.

[58] U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, vol. 1, no. 3, pp. 244–256, 1972. DOI: 10.1016/S0146-664X(72)80017-0.

[59] D. P. Fan, J. Zhang, G. Xu, M. M. Cheng, L. Shao. Salient objects in clutter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to be published. DOI: 10.1109/TPAMI.2022.3166451.

[60] D. P. Fan, Z. Lin, Z. Zhang, M. L. Zhu, M. M. Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2021. DOI: 10.1109/TNNLS.2020.2996406.

[61] X. L. Wang, R. Girshick, A. Gupta, K. M. He. Non-local neural networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, USA, pp. 7794–7803, 2018. DOI: 10.1109/CVPR.2018.00813.

[62] Y. C. Gu, L. J. Wang, Z. Q. Wang, Y. Liu, M. M. Cheng, S. P. Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of AAAI Conference on Artificial Intelligence*, Palo Alto, USA, vol. 34, pp. 10869–10876, 2020. DOI: 10.1609/aaai.v34i07.6718.

[63] L. T. Guo, J. Liu, X. X. Zhu, P. Yao, S. C. Lu, H. Q. Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, USA, pp. 10327–10336, 2020. DOI: 10.1109/CVPR42600.2020.01034.

[64] J. L. Ba, J. R. Kiros, G. E. Hinton. Layer normalization. [Online], Available: https://arxiv.org/abs/1607.06450, 2016.

[65] S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, P. Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021. DOI: 10.1109/TPAMI.2019.2938758.

[66] S. T. Liu, D. Huang, Y. H. Wang. Receptive field block net for accurate and fast object detection. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 404–419, 2018. DOI: 10.1007/978-3-030-01252-6_24.

[67] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[68] P. Krähenbühl, V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, ACM, Granada, Spain, pp. 109–117, 2011. DOI: 10.5555/2986459.2986472.

[69] X. K. Lu, W. G. Wang, C. Ma, J. B. Shen, L. Shao, F. Porikli. See more, know more: Unsupervised video object segmentation with co-attention Siamese networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 3618–3627, 2019. DOI: 10.1109/CVPR.2019.00374.

[70] T. F. Zhou, J. W. Li, S. Z. Wang, R. Tao, J. B. Shen. MATNet: Motion-attentive transition network for zero-

shot video object segmentation. *IEEE Transactions on Image Processing*, vol. 29, pp. 8326–8338, 2020. DOI: 10.1109/TIP.2020.3013162.

[71] R. T. Liu, Z. R. Wu, S. X. Yu, S. Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *Proceedings of the Advances in Neural Information Processing Systems*, online, pp. 13137–13152, 2021.

[72] M. Zhang, J. Liu, Y. F. Wang, Y. Piao, S. Y. Yao, W. Ji, J. Li, H. C. Lu, Z. X. Luo. Dynamic context-sensitive filtering network for video salient object detection. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 1533–1543, 2021. DOI: 10.1109/ICCV48922.2021.00158.

[73] G. P. Ji, K. R. Fu, Z. Wu, D. P. Fan, J. B. Shen, L. Shao. Full-duplex strategy for video object segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, IEEE, Montreal, Canada, pp. 4902–4913, 2021. DOI: 10.1109/ICCV48922.2021.00488.

[74] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk. Frequency-tuned salient region detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 1597–1604, 2009. DOI: 10.1109/CVPR.2009.5206596.

[75] D. P. Fan, G. P. Ji, X. B. Qin, M. M. Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, vol. 51, no. 9, pp. 1475–1489, 2021. DOI: 10.1360/SSI-2020-0370. (in Chinese)

[76] M. M. Cheng, D. P. Fan. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2622–2638, 2021. DOI: 10.1007/s11263-021-01490-8.

[77] R. Margolin, L. Zelnik-Manor, A. Tal. How to evaluate foreground maps?" In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 248–255, 2014. DOI: 10.1109/CVPR.2014.39.

[78] A. Borji, M. M. Cheng, H. Z. Jiang, J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015. DOI: 10.1109/TIP.2015.2487833.

[79] D. P. Fan, M. M. Cheng, Y. Liu, T. Li, A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of IEEE International Conference on Computer Vision*, Venice, Italy, pp. 4558–4567, 2017. DOI: 10.1109/ICCV.2017.487.

[80] D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 698–704, 2018. DOI: 10.24963/ijcai.2018/97.

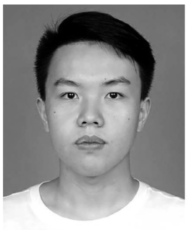
[81] D. P. Fan, G. P. Ji, M. M. Cheng, L. Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to be published. DOI: 10.1109/TPAMI.2021.3085766.

[82] G. P. Ji, D. P. Fan, Y. C. Chou, D. Dai, A. Liniger, L. Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, to be published. DOI: 10.1007/S11633-022-1365-9

[83] X. Q. Guo, J. Liu, Y. X. Yuan. Semantic-oriented labeled-to-unlabeled distribution translation for image segmentation. *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 434–445, 2022. DOI: 10.1109/TMI.2021.3114329.

[84] I. B. Senkyire, Z. Liu. Supervised and semi-supervised methods for abdominal organ segmentation: A review. *International Journal of Automation and Computing*, vol. 18, no. 6, pp. 887–914, 2021. DOI: 10.1007/s11633-021-1313-0.

[85] K. Zou, X. D. Yuan, X. J. Shen, M. Wang, H. Z. Fu. TbraTS: Trusted brain tumor segmentation. [Online], Available: https://arxiv.org/abs/2206.09309.

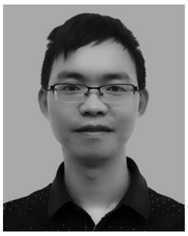


Ge-Peng Ji received the M. Sc. degree in communication and information systems from Wuhan University, China in 2021. He is currently a Ph. D. student at Australian National University, supervised by Professor Nick Barnes, majoring in Engineering and Computer Science. He has published about 10 peer-reviewed journal and conference papers. In 2021, he received the Student Travel Award from Medical Image Computing and Computer-Assisted Intervention Society.

His research interests lie in computer vision, especially in a variety of dense prediction tasks, such as video analysis, medical image segmentation, camouflaged object segmentation, and saliency detection.

E-mail: gepengai.ji@gmail.com

ORCID iD: 0000-0001-7092-2877



Guobao Xiao received the Ph. D. degree in computer science and technology from Xiamen University, China in 2016. From 2016–2018, he was a postdoctoral fellow at School of Aerospace Engineering, Xiamen University, China. He is currently a professor at Minjiang University, China. He has published over 50 papers in international journals and conferences, including TPAMI/TIP/TITS/TIE/TMM, IJCV, PR, ICCV, ECCV, etc. He has been awarded the best Ph. D. thesis in Fujian Province and the best Ph. D. thesis award in China Society of Image and Graphics (a total of ten winners in China). He also served on the program committee (PC) of CVPR, ICCV, ECCV, etc.

His research interests include machine learning, computer vision and pattern recognition.

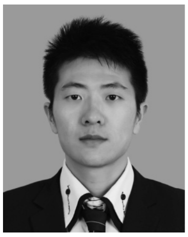
E-mail: x-gb@163.com

ORCID iD: 0000-0003-2928-8100

Yu-Cheng Chou received the B. Sc. degree in software engineering from School of Computer Science, Wuhan University, China in 2022. He is currently a visiting student at Johns Hopkins University, supervised by Zongwei Zhou and Prof. Alan Yuille.

His research interests include medical imaging, causality, and computer vision, especially developing novel methodologies to detect lesions accurately and exploring explainability through causality for computer-aided diagnosis and surgery.

E-mail: johnson111788@gmail.com
ORCID iD: 0000-0002-9334-2899



Deng-Ping Fan received the Ph. D. degree from Nankai University, China in 2019. He joined the Inception Institute of Artificial Intelligence (IIAI), UAE in 2019. He is a Postdoctoral Researcher, working with Prof. Luc Van Gool in Computer Vision Laboratory, ETH Zürich, Switzerland. He has published about 50 top journal and conference papers such as TPAMI, IJCV, TIP, TNNLS, TMI, CVPR, ICCV, ECCV, IJCAI, etc. He won the Best Paper Finalist Award at *IEEE CVPR* 2019, and the Best Paper Award Nominee at *IEEE CVPR* 2020. He was recognized as the CVPR 2019 outstanding reviewer with a special mention award, the CVPR 2020 outstanding reviewer, the ECCV 2020 high-quality reviewer, and the CVPR 2021 outstanding reviewer. He served as a program committee board (PCB) member of IJCAI 2022–2024, a senior program committee (SPC) member of IJCAI 2021, a program committee member (PC) of CAD&CG 2021, a committee member of China Society of Image and Graphics (CSIG), area chair in NeurIPS 2021 Datasets and Benchmarks Track, area chair in MICCAI2020 Workshop.

His research interests include computer vision, deep learning, and saliency detection.

E-mail: dengpfan@gmail.com
ORCID iD: 0000-0002-5245-7518 (Corresponding author)

Kai Zhao received the B. Sc. and M. Sc. degrees from Shanghai University, China in 2014 and 2017, respectively, and the Ph. D. degree from College of Computer Science, Nankai University, China in 2020. He is currently a postdoctoral researcher at University of California, USA. He has over 10 peer-reviewed publications in computer vision and machine learning-related areas, including TPAMI, TIP, NeurIPS, ICCV, CVPR, ECCV and IJCAI.

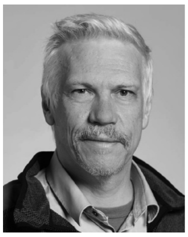
His research interests include computer vision and machine intelligence.

E-mail: kz@kaizhao.net
ORCID iD: 0000-0002-2496-0829

Geng Chen received the Ph. D. degree from Northwestern Polytechnical University, China in 2016. He was a research scientist at the Inception Institute of Artificial Intelligence, UAE from 2019 to 2021, and a postdoctoral research associate at the University of North Carolina at Chapel Hill, USA from 2016 to 2019. He is a professor at Northwestern Polytechnical University, China. He has published over 60 papers in peer-reviewed international conference proceedings and journals.

His research interests include medical image analysis and computer vision.

E-mail: geng.chen.cs@gmail.com
ORCID iD: 0000-0001-8350-6581



Luc Van Gool received the B. Eng. degree in electromechanical engineering from the Katholieke Universiteit Leuven in 1981. Currently, he is a professor at the Katholieke Universiteit Leuven in Belgium and the ETH Zürich, Switzerland. He leads computer vision research at both places and also teaches at both 5. He has been a program committee member of several major computer vision conferences. He received several Best Paper awards, won a David Marr Prize and a Koenderink Award, and was nominated Distinguished Researcher by the IEEE Computer Science committee. He is a co-founder of 10 spin-off companies.

His research interests include 3D reconstruction and modeling, object recognition, tracking, gesture analysis, and a combination of those.

E-mail: vangool@vision.ee.ethz.ch
ORCID iD: 0000-0002-3445-5711