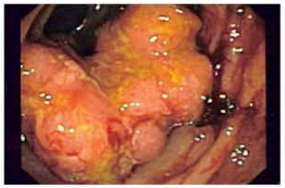
Deep Learning and Hand-crafted Feature Based Approaches for Polyp Detection in Medical Videos

Konstantin Pogorelov, Olga Ostroukhova, Mattis Jeppsson, Håvard Espeland,
Carsten Griwodz, Thomas de Lange, Dag Johansen, Michael Riegler and Pål Halvorsen

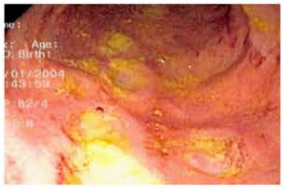
Gastrointestinal (GI) tract diseases

- Many types of diseases can potentially affect the human GI tract

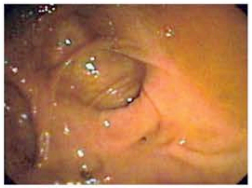
(a) Colon polyp



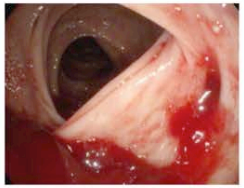
(b) Colorectal cancer



(c) Crohn's disease



(d) Diverticulosis



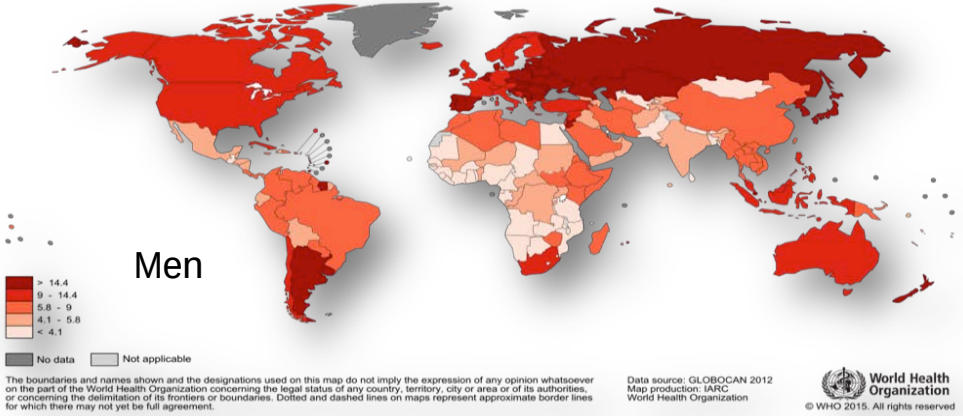
(e) Bleeding

(f) Anastomosis

  - about 2.8 millions new luminal GI cancers (esophagus, stomach, colorectal) are detected yearly
  - the mortality is about 65%

- Screening of the GI tract using different types of endoscopy...
  - is costly (colonoscopy according to NY Times: $1100/patient, $10 billion dollars)
  - consumes valuable medical personnel time (1-2 hours)
  - does not scale to large populations
  - is intrusive to the patient
  - ...

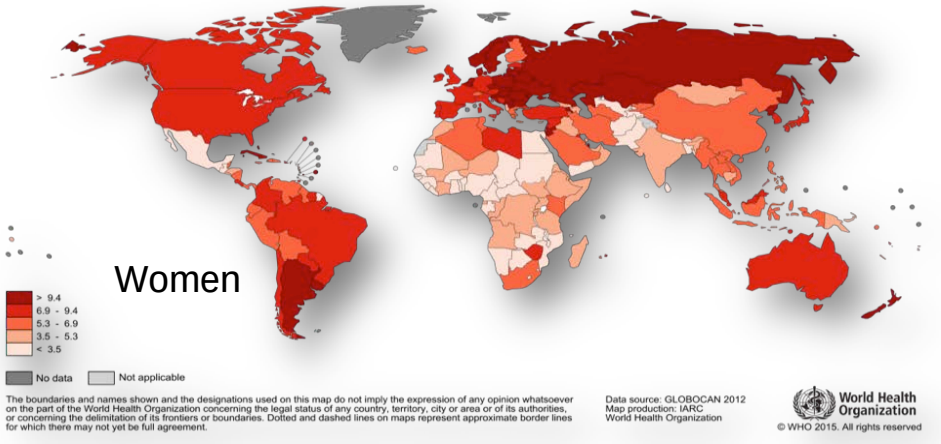
WHO: Colorectal Cancer Mortality 2015



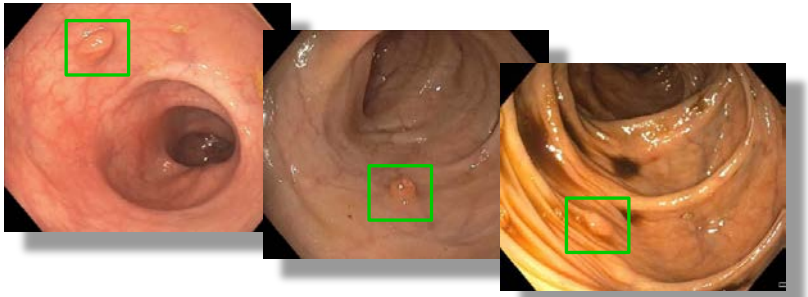
Colorectal cancer is the **third most common cause of cancer** mortality for both women and men, and it is a condition where **early detection is important for survival**, i.e., a 5-year survival probability of going from a low 10-30% if detected in later stages to a high 90% survival probability in early stages.

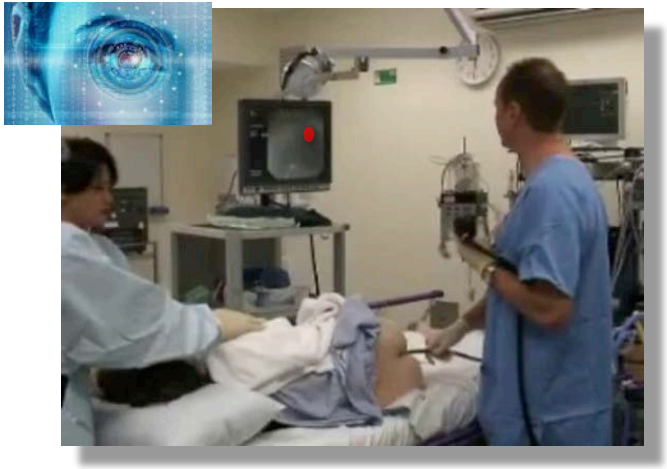
Related to the cancer example, on average **20% of polyps (possible predecessors of cancer) are missed** or incompletely removed. The risk of getting cancer largely depend on the endoscopists ability to detect and remove polyps.

A **1% increase in detection can decrease the risk of cancer with 3%**.



Standard endoscopy: Live Polyp Detection



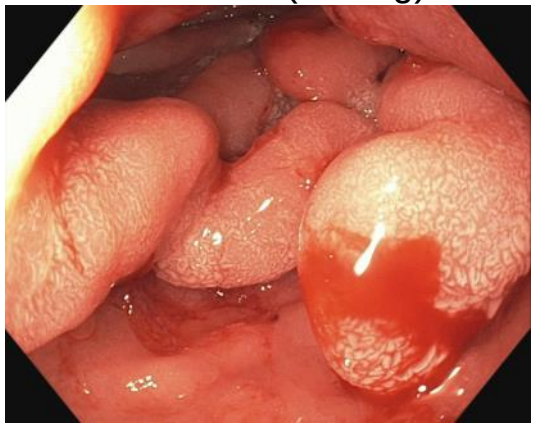


- A **polyp** is an abnormal growth of tissue attached to the underlying mucosa
- Detection accuracy depends on experience and skills
  - average miss rates of approx. **20%**
  - large inter- and intra-variations
  - should reach a high (>**85%**) accuracy threshold to be acceptable
- Current technology may potentially enable **automated algorithmic** assisted examinations
- Introduce a digital “third eye”
  (with high accuracy and real-time processing)

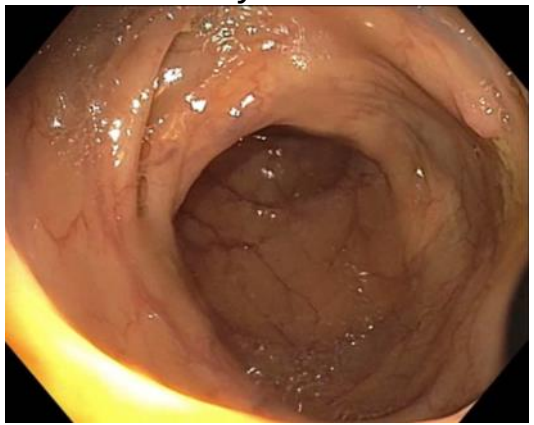
Polyps detection: challenges

Flat / same color

Too close (too big)



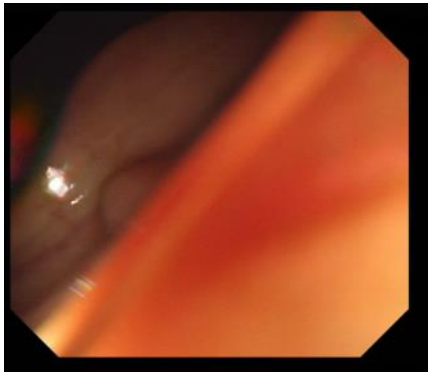
Partly hidden



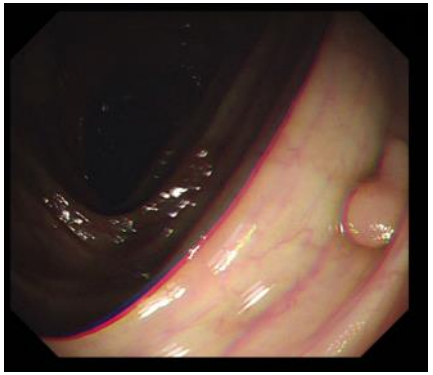
Aux. data / nav. box

Lens contamination

Blur / motion blur



Colors shift



State-of-The-Art: Some Example Detection Systems

Publication/ System	What/ Detection Types	Recall/ Sensitivity	Precision	Specificity	Accuracy	FPS	Dataset Size
Wang et al. [40]	polyp/edge, texture	97.7%*	–	95.7%	–	10	1.8m frames
Wang et al. [39]	polyp/shape,color,texture	81.4%	–	–	–	0.14	1,513 images
Mamonov et al. [19]	polyp/shape	47%	–	90%	–	–	18,738 frames
Hwang et al. [14]	polyp/shape	96%	83%	–	–	15	8,621 frames
Li and Meng [17]	tumor/textural pattern	88.6%	–	96.3%	92.4%	–	–
Zhou et al. [42]	polyp/intensity	75%	–	95.92%	90.8%	–	–
Alexandre et al. [4]	polyp/color pattern	93.7%	–	76.9%	–	–	35 images
Kang et al. [16]	polyp/shape,color	–	–	–	–	1	–
Cheng et al. [9]	polyp/texture,color	86.2%	–	–	–	0.08	74 images
Ameling et al. [5]	polyp/texture	AUC=95%	–	–	–	–	1,736 images

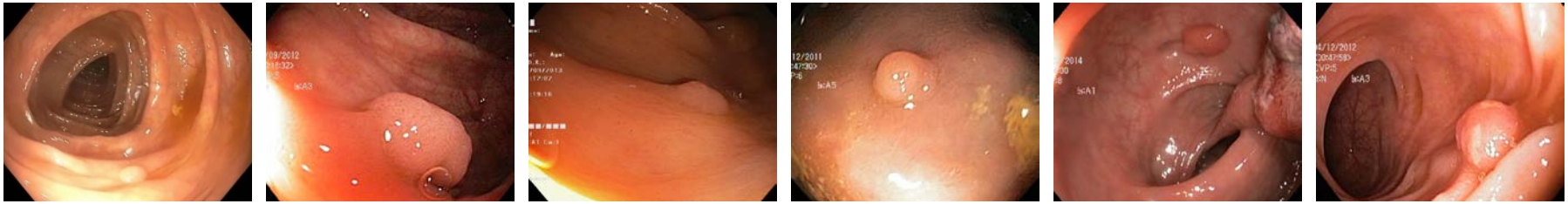
Good, but can we do better?
We have earlier tried solutions based using Global features (GF) and Convolutional neural network (CNN)

Previous Study: EIR on ASU Mayo Polyp Dataset

- Mayo dataset (18781 images)
- Handcrafted global features:
  - recall **98.50%**, precision **93.88%**, fps **~300**
- Deep-learning features using CNN:
  - Modified Inception v3: recall **95.86%**, precision **80.78%**, fps: **~30**
  - Inception v3 + WEKA: recall: **88.87%**, precision: **89.16%**, fps: **~30**

Similar numbers on other datasets!

Very good numbers, but generalizable?
Will it work in a real clinical setting
(different doctors, different equipment, ...)?



Mixing Datasets

- Combination of multiple different datasets

- Datasets from traditional colonoscopy – differently skewed
  - GIANA 2017 challenge (**CVC**):
    - 12,922 frames (10,993 positive samples + 1,929 negative samples)
  - **Kvasir** + **Nerthus**:
    - 7,350 frames (1,000 positive samples + 6,350 negative, non-polyp samples)

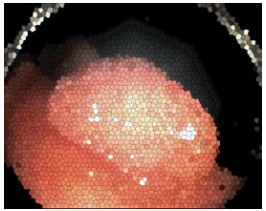
Dataset	Training	Test	# Frames	# Polyp frames	# Normal frames
CVC-356	X	X	1,706	356	1,350
CVC-612	X	X	1.962	612	1,350
CVC-968	X	X	2.318	968	1,350
CVC-12k	-	X	11,954	10,025	1,929
Kvasir	-	X	6,000	1,000	5,000
Nerthus	X	-	1,350	-	1,350

meaning non-polyp frames,
but not normal mucosa – other findings

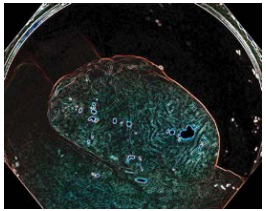
Hand-crafted Global Features

- Features extraction using open-source LIRE (Lucene Image Retrieval)
- LIRE image feature descriptors JCD and Tamura are the best choice
- Search-based classification using the Logistic Model Tree (LMT) classifier
- Combination of features using early fusion

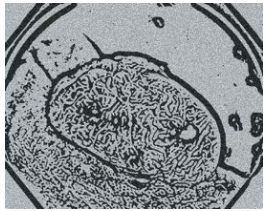
Original polyp



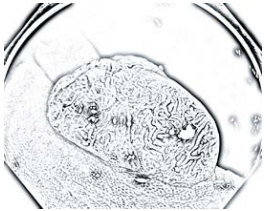
Color feature



Edge and color



Texture



Edge

Table I. Leave-one-out cross-evaluation combined for all supported features.

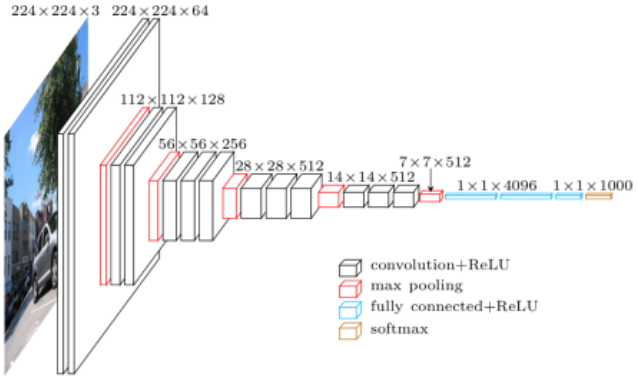
Feature	True Pos.	True Neg.	False Pos.	False Neg.	Precision	Recall	F1 score
JointHist.	3,369	13,826	1,085	511	0.7563	0.8682	0.8084
JpegCoefficientHist.	3,224	13,772	1,139	656	0.7389	0.8309	0.7822
Tamura	3,392	13,861	1,050	488	0.7636	0.8742	0.8151
FuzzyOpponentHist.	3,341	13,552	1,359	539	0.7108	0.8610	0.7787
SimpleColorHist.	2,736	13,563	1,348	1,144	0.6699	0.7051	0.6870
JCD	3,556	13,777	1,134	324	0.7582	0.9164	0.8298
FuzzyColorHist.	2,708	13,243	1,668	1,172	0.6188	0.6979	0.6560
RotationInvariantLBP	3,479	13,829	1,082	401	0.7627	0.8966	0.8243
FCTH	2,846	13,671	1,240	1,034	0.6965	0.7335	0.7145
LocalBinaryPatterns-AndOpponent	2,412	13,349	1,562	1,468	0.6069	0.6216	0.6142
PHOG	2,879	13,806	1,105	1,001	0.7226	0.7420	0.7321
RankAndOpponent	2,527	13,553	1,358	1,353	0.6504	0.6512	0.6508
ColorLayout	2,702	14,018	893	1,178	0.7515	0.6963	0.7229
CEDD	3,705	13796	1,115	175	0.7686	0.9548	0.8517
Gabor	1,849	10,643	4,268	2,031	0.3022	0.4765	0.3699
OpponentHist.	2,246	14,157	754	1,634	0.7486	0.5788	0.6529
EdgeHist.	3,548	13,737	1,174	332	0.7513	0.9144	0.8249
ScalableColor	3,231	13,684	1,227	649	0.7247	0.8327	0.7750
Late Fusion	3,710	13,894	1,017	170	0.7848	0.9561	0.8620

Table II. Top 20 results for feature combinations using two image features for the video wp_61, sorted by F1 score.

Feature combinations	True Pos.	True Neg.	False Pos.	False Neg.	Precision	Recall	F1 score
Rot.Inv.LBP/Tamura	162	22	153	0	0.5142	1	0.6792
PHOG/Tamura	161	23	152	1	0.5143	0.9938	0.6778
JpegCoeff.Hist./Tamura	162	21	154	0	0.5126	1	0.6778
Gabor/Tamura	162	20	155	0	0.5110	1	0.6764
FuzzyColorHist./Tamura	162	18	157	0	0.5078	1	0.6735
FuzzyOpp.Hist./FuzzyColorHist.	160	17	158	2	0.5031	0.9876	0.6666
JCD/Opp.Hist.	135	67	108	27	0.5555	0.8333	0.6666
JointHist./JpegCoeff.Hist.	162	12	163	0	0.4984	1	0.6652
ColorLayout /FuzzyColorHist.	162	11	164	0	0.4969	1	0.6639
FuzzyColorHist./JointHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./JointHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./SimpleColorHist.	162	11	164	0	0.4969	1	0.6639
JointHist./Rotat.Inv.LBP	162	11	164	0	0.4969	1	0.6639
JointHist./SimpleColorHist.	162	11	164	0	0.4969	1	0.6639
FuzzyOpp.Hist./Gabor	161	13	162	1	0.4984	0.9938	0.6639
JCD/JpegCoeff.Hist.	161	13	162	1	0.4984	0.9938	0.6639
CEDD/FuzzyColorHist.	159	17	158	3	0.5015	0.9814	0.6638
JpegCoeff.Hist./Rot.Inv.LBP	152	31	144	10	0.5135	0.9382	0.6637
JCD/Tamura	162	10	165	0	0.4954	1	0.6625
CEDD/Tamura	162	10	165	0	0.4954	1	0.6625

Deep Learning (CNN) Features

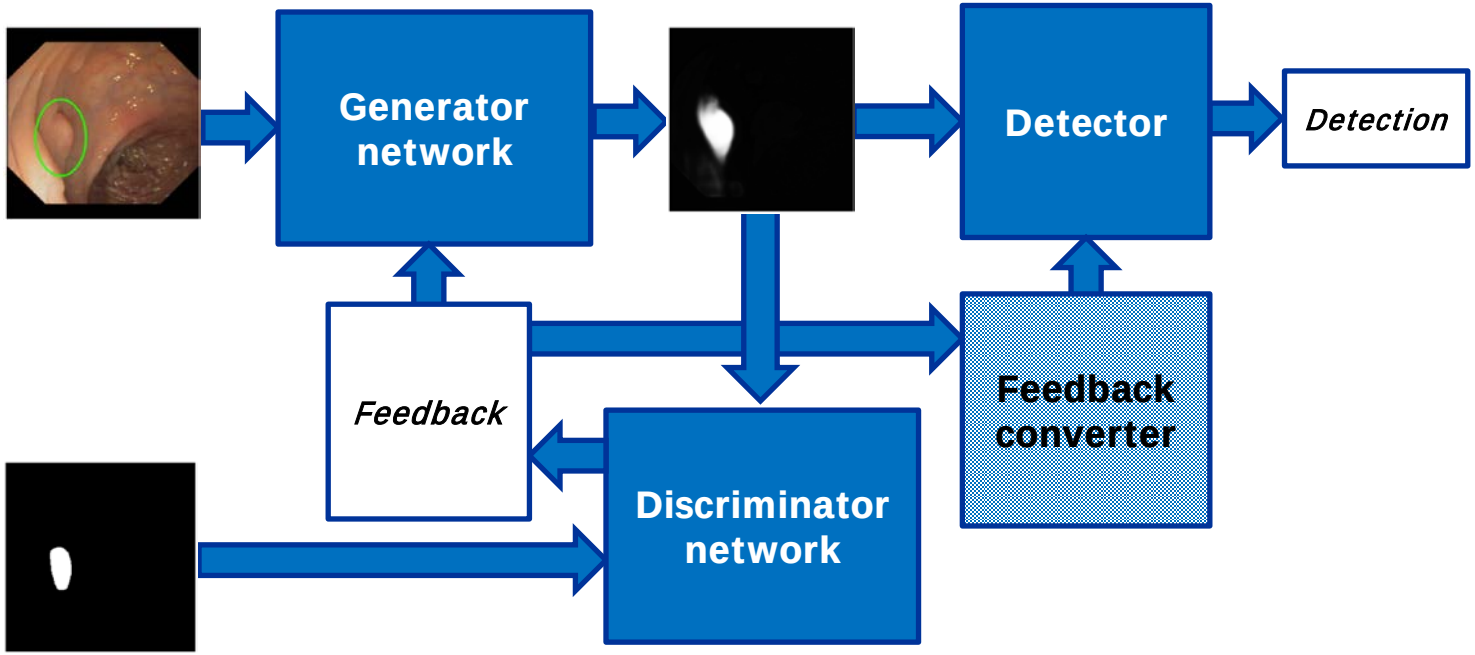
- Transfer learning (TensorFlow)
  - using Xception, VGG19 and ResNet50
  - trained on Imagenet
  - freeze and train base layers
  - tune entire network (Bayesian optimization)



- Region of interest detection (YOLO)
  - detect objects within a frame
  - trained from scratch using ground truth masks in CVC-968

Generative adversarial network (GAN)

- **Generator** and **Discriminator networks** and a threshold **activation detector**
- Provide a pixel-perfect detection map – do not miss small spots
- Data augmentation (flip + rotation giving 35 new images per image)



Detection performance

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
Kvasir	GAN-356	CVC-356	0.715	0.751	0.940	0.909	0.732	0.677
	GAN-612	CVC-612	0.595	0.803	0.891	0.876	0.684	0.619
	GAN-968	CVC-968	0.736	0.746	0.946	0.913	0.741	0.689
	GF-D-356	CVC-356	0.171	0.109	0.894	0.763	0.133	0.004
	GF-D-612	CVC-612	0.270	0.318	0.828	0.743	0.292	0.137
	GF-D-968	CVC-968	0.225	0.859	0.409	0.484	0.357	0.208
	RT-D-Xcept-356	CVC-356	0.358	0.259	0.907	0.799	0.300	0.190
	RT-D-Xcept-612	CVC-612	0.383	0.326	0.895	0.800	0.352	0.236
	RT-D-Xcept-968	CVC-968	0.459	0.256	0.939	0.825	0.328	0.251
	RT-D-VGG19-356	CVC-356	0.181	0.333	0.777	0.720	0.235	0.087
	RT-D-VGG19-612	CVC-612	0.213	0.583	0.682	0.669	0.313	0.186
	RT-D-VGG19-968	CVC-968	0.231	0.320	0.842	0.774	0.268	0.142
	RT-D-ResNe-356	CVC-356	0.236	0.178	0.885	0.767	0.203	0.070
	RT-D-ResNe-612	CVC-612	0.321	0.507	0.785	0.739	0.393	0.247
	RT-D-ResNe-968	CVC-968	0.248	0.877	0.469	0.537	0.387	0.262
	YOLO-968	CVC-968	0.530	0.559	0.901	0.844	0.544	0.450

Test set	Run	Training set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-12k	GAN-356	CVC-356	0.967	0.624	0.888	0.667	0.758	0.378
	GAN-612	CVC-612	0.934	0.609	0.778	0.636	0.737	0.286
	GAN-968	CVC-968	0.906	0.912	0.510	0.847	0.909	0.428
	GF-D-356	CVC-356	0.829	0.909	0.030	0.767	0.867	-0.081
	GF-D-612	CVC-612	0.809	0.383	0.530	0.407	0.520	-0.064
	GF-D-968	CVC-968	0.835	0.854	0.125	0.737	0.845	-0.020
	RT-D-Xcept-356	CVC-356	0.913	0.624	0.693	0.636	0.742	0.236
	RT-D-Xcept-612	CVC-612	0.876	0.740	0.457	0.694	0.802	0.160
	RT-D-Xcept-968	CVC-968	0.899	0.690	0.600	0.676	0.781	0.224
	RT-D-VGG19-356	CVC-356	0.257	0.292	0.874	0.799	0.273	0.158
	RT-D-VGG19-612	CVC-612	0.266	0.489	0.799	0.759	0.344	0.228
	RT-D-VGG19-968	CVC-968	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-356	CVC-356	0.723	0.003	0.999	0.871	0.006	0.044
	RT-D-ResNe-612	CVC-612	0.232	0.406	0.800	0.750	0.295	0.166
	RT-D-ResNe-968	CVC-968	0.870	0.303	0.766	0.378	0.450	0.057
	YOLO-968	CVC-968	0.932	0.641	0.757	0.660	0.759	0.296

approx. 1.5 fps

approx. 100 fps

approx. 30 fps

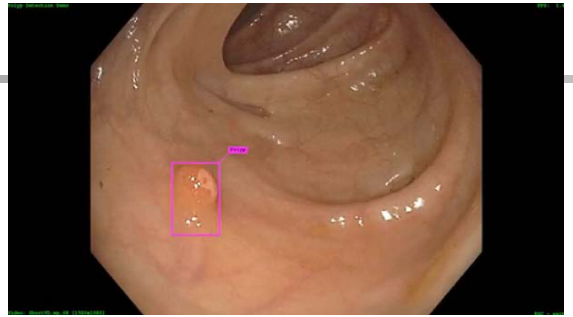
- Most of the approaches have a too low detection rate – average doctor (80%) and accuracy goal (85%)
  (but still better than the random baseline)
- Several approaches reach the accuracy goal
- GANs are in general most “accurate”: **91% and 85%**, respectively (but tooooo slooooooow)

GAN (per-pixel) localization performance

Test set	Run	Train set	PREC	SENS	SPEC	ACC	F1	MCC
CVC-612	LOC-356	CVC-356	0.819	0.619	0.984	0.946	0.706	0.684
CVC-356	LOC-612	CVC-612	0.723	0.735	0.981	0.965	0.729	0.710

(a) Input frame (b) Ground truth mask (c) Segmentation mask

Summary



- Polyp is an anomaly that is hard to detect, but many good studies exist
- Overfitting is a challenge in existing studies
  → we have mixed datasets
- Many approaches are better than the average medical expert, some even reach the goal
  (and different methods may rank different for different diseases)
- A GAN has
  - a superior detection rate
  - slow as for now, but might be ok for an offline examination (like video capsules)
  - can be improved for real-time examination support
- Ongoing and future work include
  - other diseases
  - speed improvements (GPU acceleration)

Questions?

Contact-information:

Konstantin Pogorelov
konstantin @ simula.no
https://www.simula.no/people/konstantinvpogorelov

Michael Riegler
michael @ simula.no
https://www.simula.no/people/michael

Pål Halvorsen
paalh @ simula.no
http://home.ifi.uio.no/paalh