Gastroenterologist-level detection of gastric precursor lesions and neoplasia with a deep convolutional neural network

Supplementary Figures S1-S3, Tables S1-S4 and Video S1
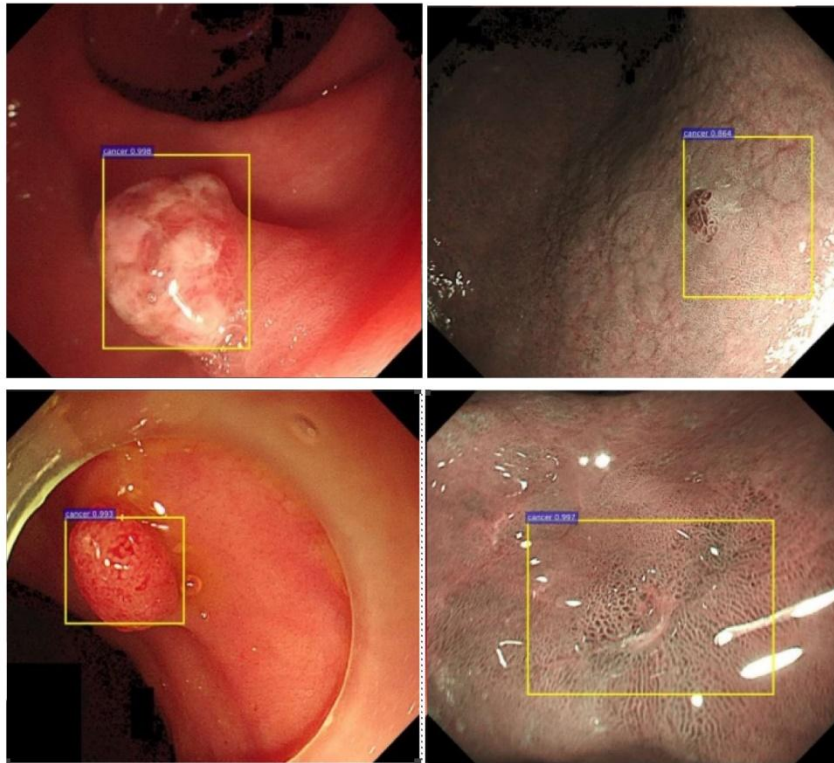
Supplementary Figures:
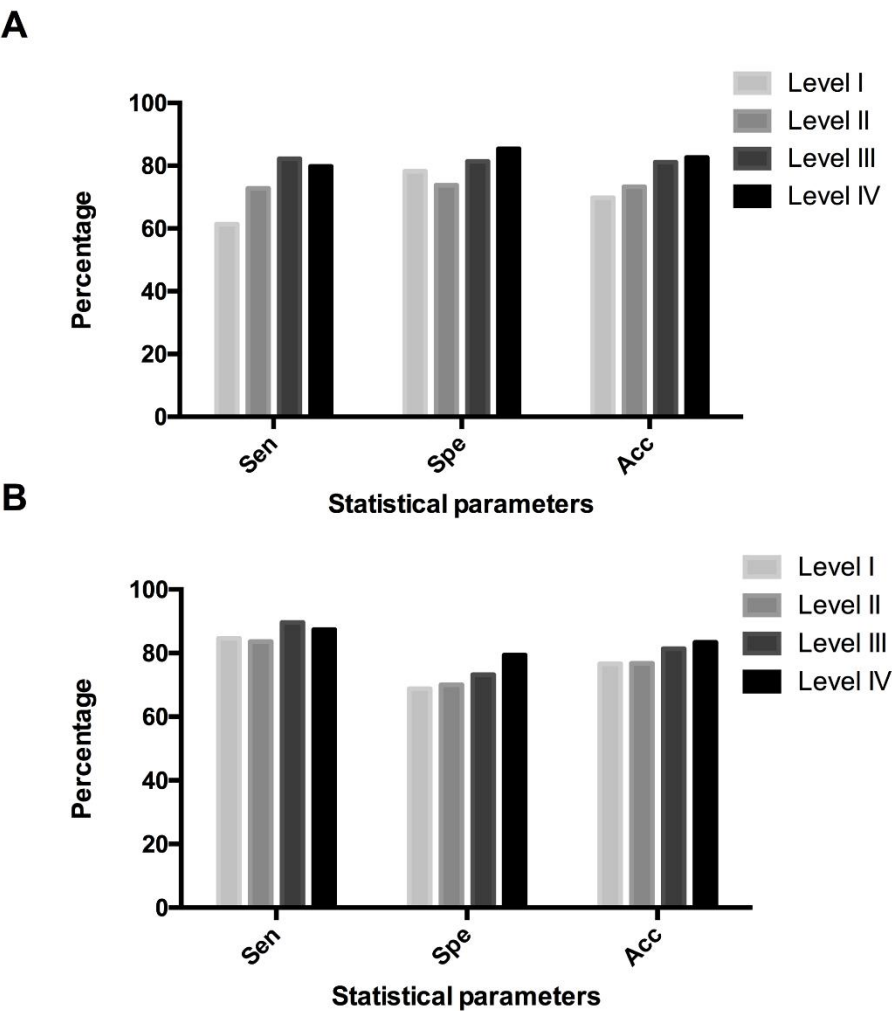
Supplementary Figure S1. Average performance of GI doctors in each level in the diagnostic test for CAG and GC. **(A)** Diagnostic reliability of doctors in different levels in the CAG test. The sensitivity of doctors from Level I to Level IV is respectively 61.4%, 72.8%, 82.2% and 79.8%. The specificity is respectively 78.2%, 73.8%, 81.4% and 85.4%. While the accuracy ranges from 69.8%, 73.3% and 81.1% to 82.6%. **(B)** Diagnostic reliability of doctors in different levels in the GC test. The sensitivity of doctors from Level I to Level IV is respectively 84.6%, 83.6%, 89.6% and 87.4%. The specificity is respectively 68.8%, 70.0%, 73.2% and 79.4%. While the accuracy ranges from 76.6%, 76.8% and 81.4% to 83.4%.
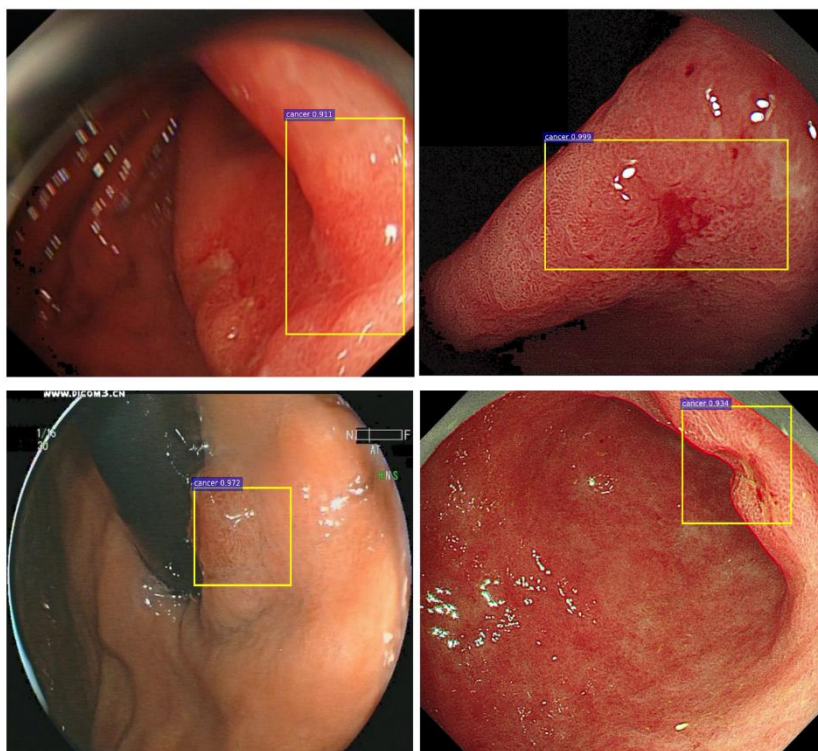
Supplementary Figure S2. Examples of unnoticeable GC lesions successfully detected by the network.

Supplementary Figure S3. Examples of gastric non-cancer lesions misdiagnosed by the network.

Supplementary Tables:

Supplementary Table S1. Detailed composition of the test pack of gastric cancer.

Composition	No. (%) of lesions
Gastric cancer (early and advanced)	50 (50.0)
Non-cancer diseases	
Benign peptic ulcers	22 (22.0)
Benign gastric polyps	14 (14,0)
GIST	5 (5.0)
Gastric heterotopia pancreas	9 (9.0)
Total	100 (100.0)

All diagnosis of lesions was proved by histopathological results.

Abbreviations: GIST, gastrointestinal stromal tumors.

Supplementary Table S2A. Relationship between diagnosis by the VGG16 model with the best accuracy rate and histological findings in CAG detection.

	Positive diagnosis by computer	Negative diagnosis by computer	Total
Positive	47	3	50
Negative	7	43	50
Total	54	46	100

Accuracy, 90.0% (90/100); sensitivity, 94.0%; and specificity, 86.0%; Positive predictive value/negative predictive value: 87.0%/93.5%.

Supplementary Table S2B. Relationship between diagnosis by the ZF model with the best accuracy rate and histological findings in GC detection.

	Positive diagnosis by computer	Negative diagnosis by computer	Total
Positive	45	5	50
Negative	25	25	50
Total	70	30	100

Accuracy, 70.0% (70/100); sensitivity, 90.0%; and specificity, 50%; Positive predictive value/negative predictive value: 64.3%/83.3%.

Supplementary Table S3. Baseline characteristics of GI doctors tested.

	CAG test No. (%)	GC test No. (%)
Gender		
Male	30 (39.0)	36 (40.4)
Female	47 (61,0)	53 (59.6)
Years of endoscopic operation		
<5 y	17 (22.0)	20 (22.4)
5-10 y	14 (18.2)	15 (16.9)
10-15 y	10 (13.0)	15 (16.9)
≥15 y	36(46.8)	39 (43.8)
Cases of gastroscopy		
≤200	20 (26.0)	20 (22.5)
201-500	10 (13.0)	14 (15.7)
501-1000	15 (19.5)	16 (18.0)
>1000	32 (41.5)	39 (43.8)
Cases of colonoscopy		
≤200	29 (37.7)	34 (38.2)
201-500	14 (18.2)	14 (15.7)
501-1000	11 (14.3)	15 (16.9)
>1000	23 (29.8)	26 (29.2)
Total	77 (100.0)	89 (100.0)

Supplementary Table S4. Intra-observer agreement of doctors in different levels regarding CAG and GC diagnosis compared with histological findings.

		Level I	Level II	Level III	Level IV
CAG diagnosis	Kappa*	0.261	0.284	0.468	0.514
	95% CI	0.244-0.278	0.264-0.305	0.439-0.498	0.506-0.522
	SE	0.009	0.01	0.015	0.004
GC diagnosis	Kappa*	0.43	0.418	0.507	0.584
	95% CI	0.416-0.444	0.399-0.438	0.500-0.513	0.577-0.592
	SE	0.007	0.01	0.003	0.004

Supplementary Video. An example of real-time detection of CAG built upon the best CNN model.