Deep Learning-Based Anatomical Site Classification for Upper Gastrointestinal Endoscopy

Qi He[1], Sophia Bano[2], Omer F. Ahmand[2], Bo Yang[3],
Xin Chen[3], Pietro Valdastri[4], Laurence B. Lovat[2],
Danail Stoyanov[2] and Siyang Zuo[1]

[1]Tianjin University, Tianjin, China

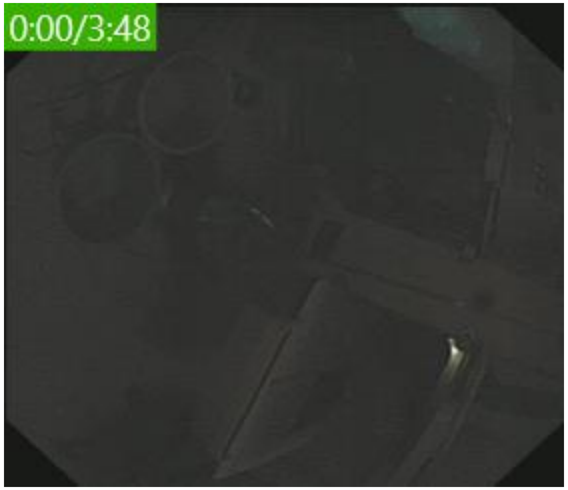
[2]University College London, WEISS, London, UK
[3]General Hospital, Tianjin Medical University, Tianjin, China

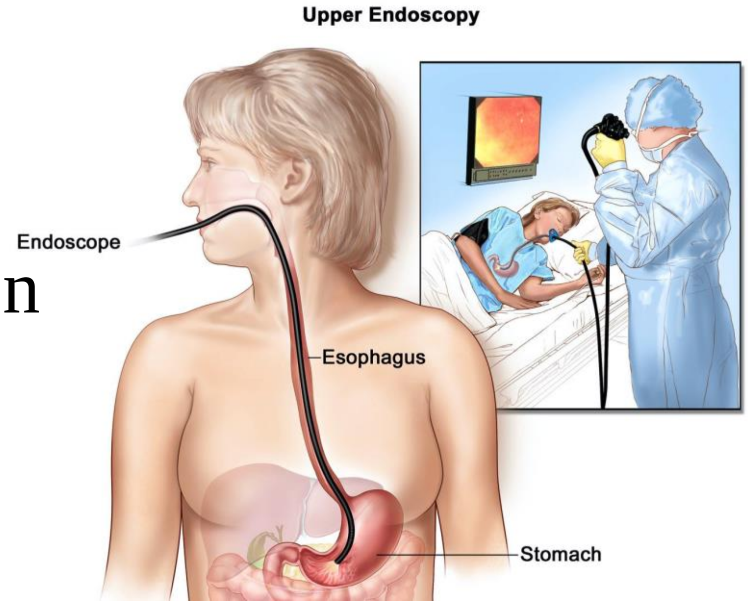
[4]University of Leeds, STORM LAB UK, Leeds, UK

Background

- Esophagogastroduodenoscopy (EGD)
  - Gold-standard
  - Widely performed
  - Potential blind spots
- Difficulties:

  Standardized photo-documentation
  - Quality indicator
  - Various guidelines
  - Time-consuming





[https://www.teresewinslow.com/]

- Need for the automatic photo-documentation method to support and efficiently improve the quality of endoscopy

Challenges

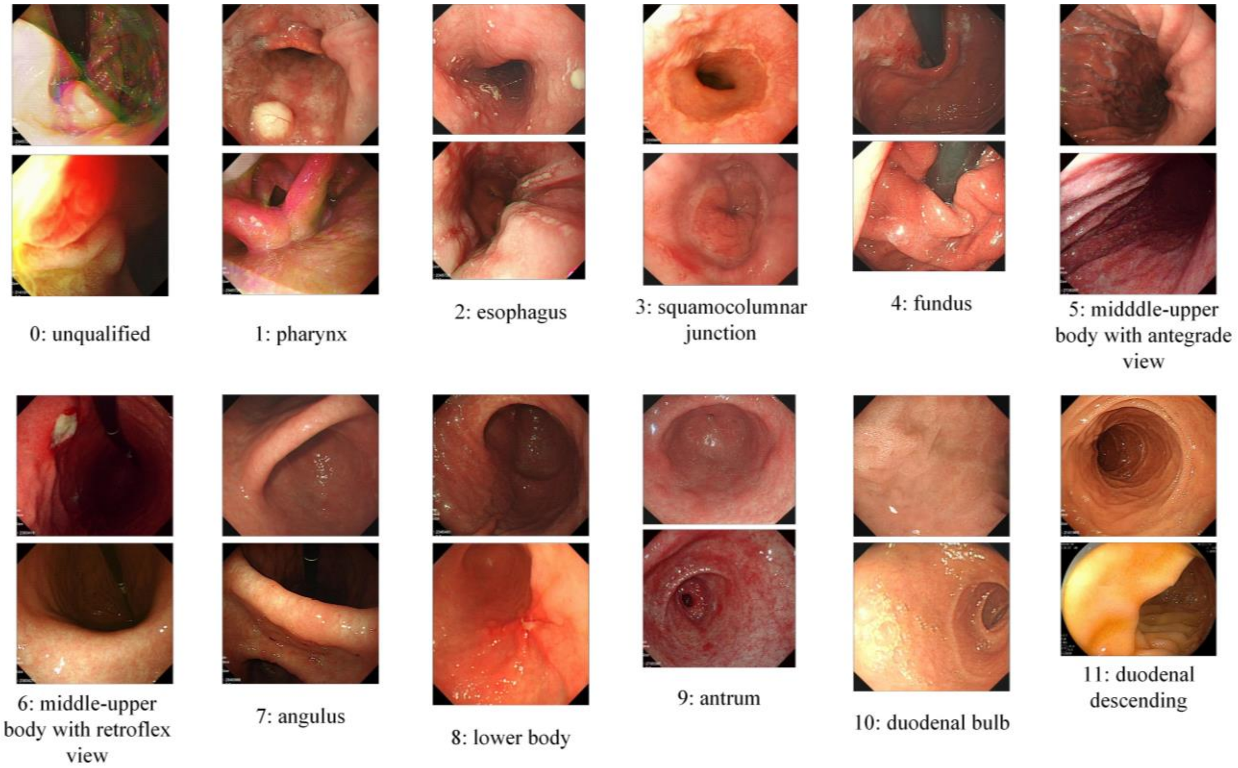
- Complete examination
  - Geographical regions with higher gastric disease incidence
  - Captured photos could construct a complete quality indicator
- Anatomical site classification
  - Easily recognized from their statics appearances
  - Cover the pre-collected image datasets as much as possible
  - Learn from a small dataset
- Need for a guideline adapted with the examination procedure and classification algorithm at the same time

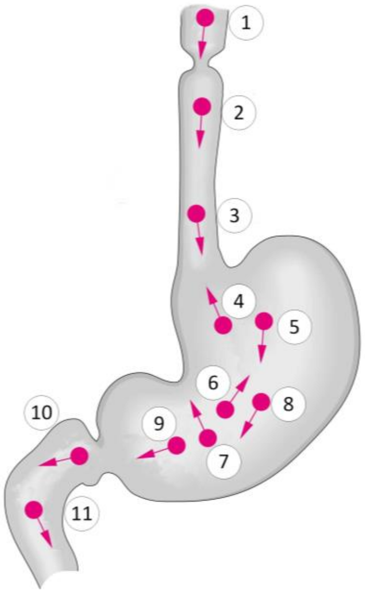
Endoscopy guidelines

- Japanese guideline [Yao, ‘13]
  - Focuses exclusively on detailed imaging of the stomach including comprehensive multiple quadrant views of each landmark
  - Not routinely clinically implemented outside of Japan
- British guideline [BSG and AUGIS, ‘17] [ESGE, ‘01]
  - Includes additional important landmarks outside of the stomach
  - Fewer images of the stomach
- Need for designing a new upper GI guideline that adapted to existing examination procedure.

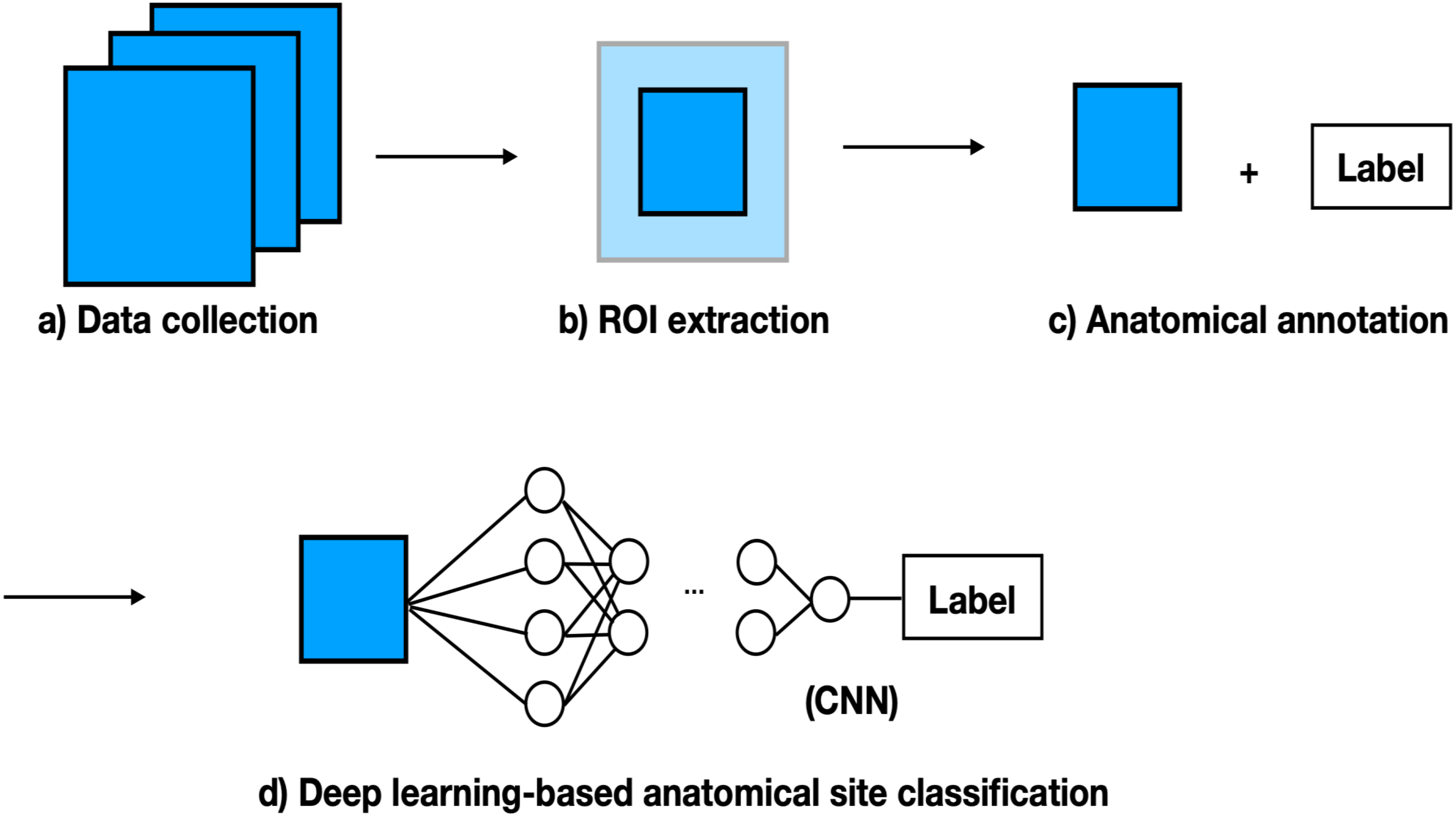
Objectives

- **Guideline**
    - Adapted to existing examination procedure
    - Robust quality indicator
    - Annotation friendly





Workflow



a) Data collection

b) ROI extraction

c) Anatomical annotation

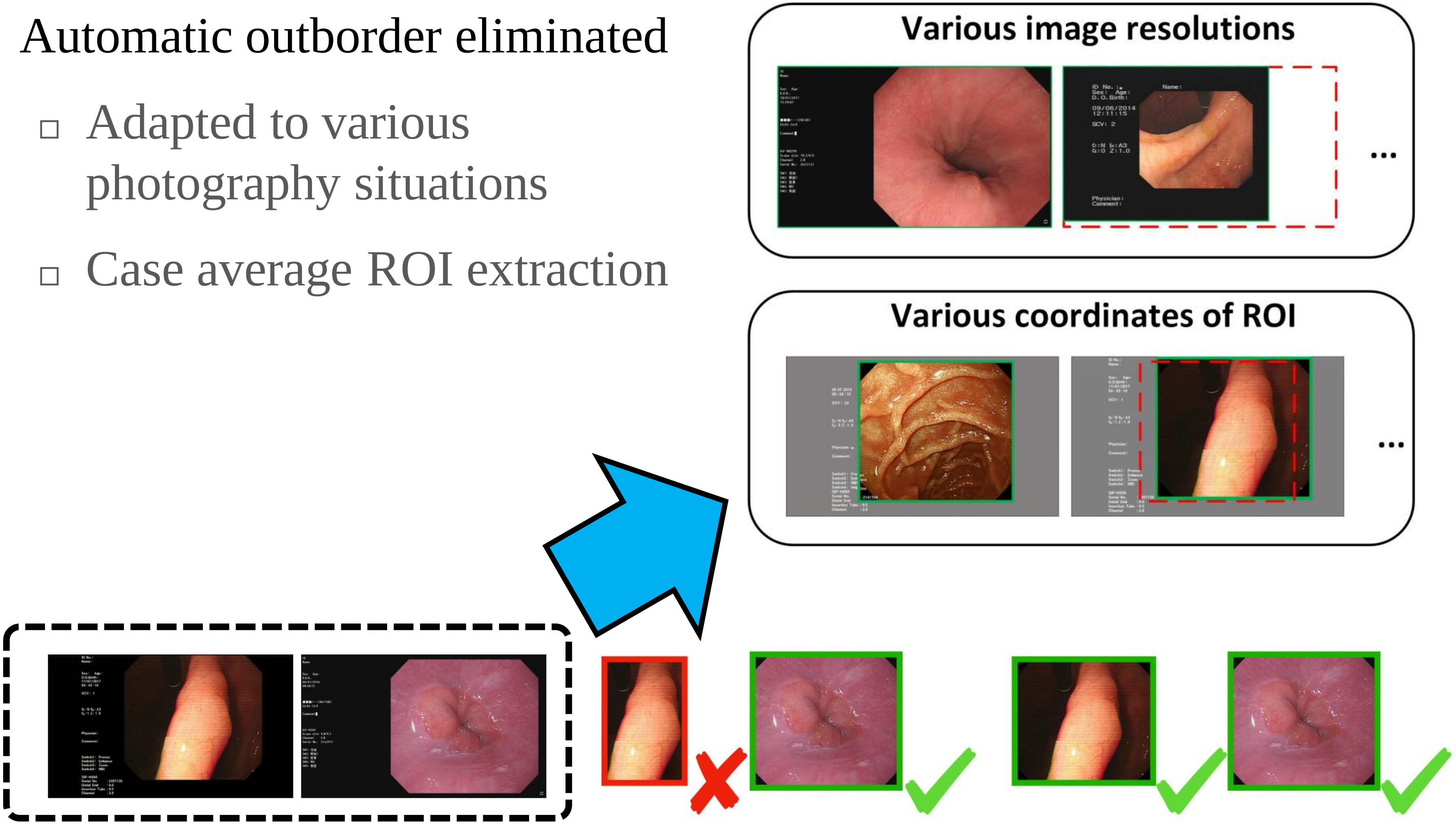
d) Deep learning-based anatomical site classification

Design of data collection

- Dataset before preprocessing
    - Image resolution: 768 x 578, 1024 x 600…
    - Imaging mode: WL, LCI, NBI…
    - Dataset size: 229 cases including 5661 images
- Dataset after preprocessing
    - Imaging mode: WL, LCI
    - Dataset size: 211 cases including 3704 images

Design of ROI extraction

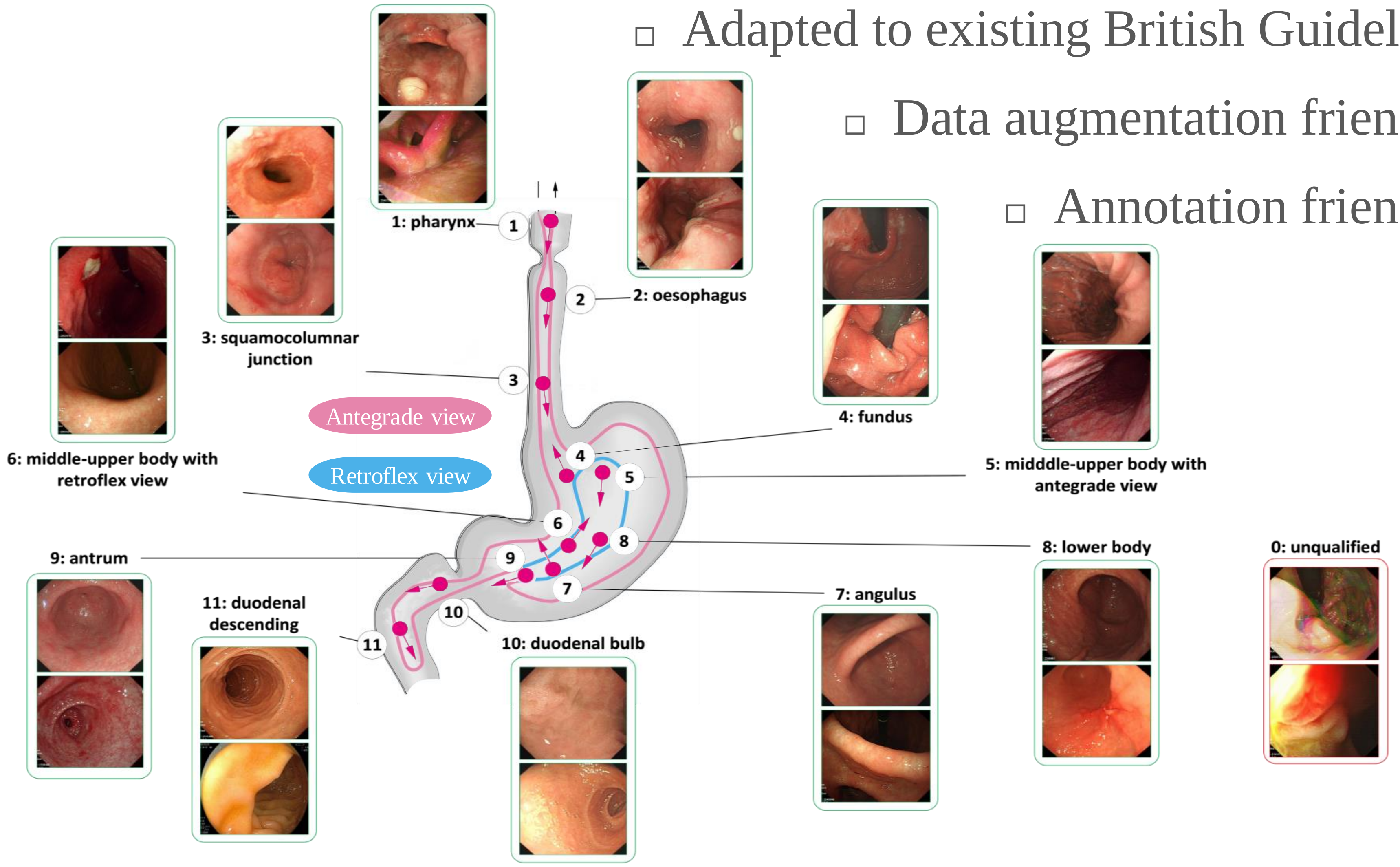
- Automatic outborder eliminated
  - Adapted to various photography situations
  - Case average ROI extraction



a) Original images b) ROIs by threshold c) Our case average ROIs

Design of Anatomical annotation

- Anatomical classification guideline



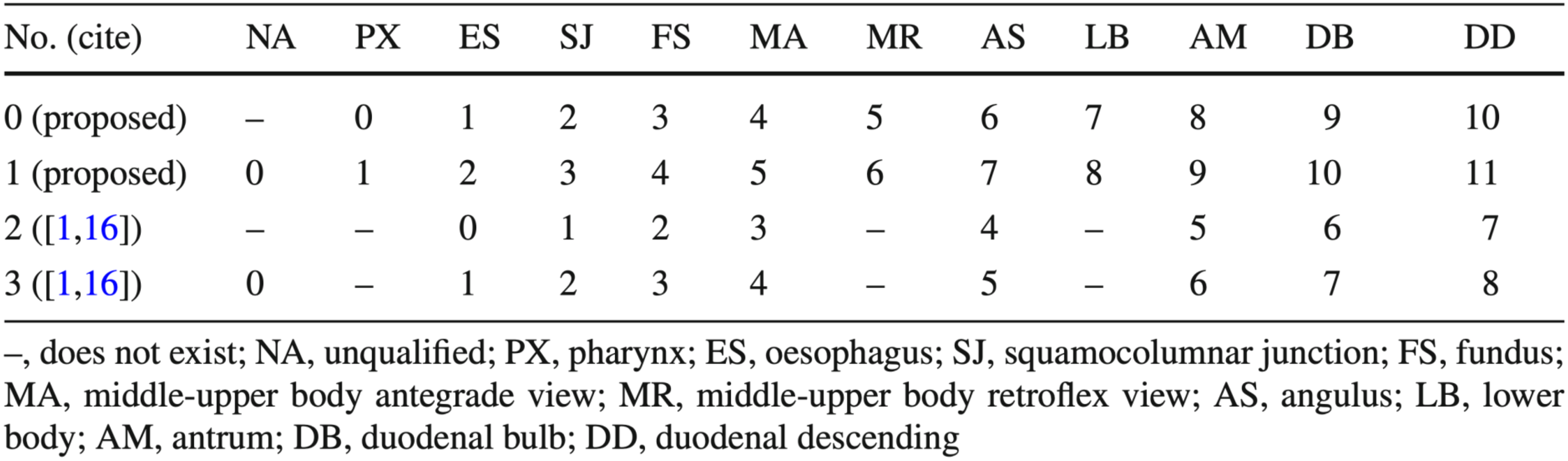
□ Adapted to existing British Guideline

□ Data augmentation friendly

□ Annotation friendly

Experimental Design

- Materials
  - Four different forms of datasets
  - Five-fold cross-validation



No. (cite)	NA	PX	ES	SJ	FS	MA	MR	AS	LB	AM	DB	DD
0 (proposed)	–	0	1	2	3	4	5	6	7	8	9	10
1 (proposed)	0	1	2	3	4	5	6	7	8	9	10	11
2 ([1,16])	–	–	0	1	2	3	–	4	–	5	6	7
3 ([1,16])	0	–	1	2	3	4	–	5	–	6	7	8

–, does not exist; NA, unqualified; PX, pharynx; ES, oesophagus; SJ, squamocolumnar junction; FS, fundus; MA, middle-upper body antegrade view; MR, middle-upper body retroflex view; AS, angulus; LB, lower body; AM, antrum; DB, duodenal bulb; DD, duodenal descending

Experimental Design

- Evaluation metrics and model implementation
    - The overall accuracy (models):

$$\mathrm{rate}_{oa}(Y, f(X)) = \frac{\mathrm{sum}(\mathrm{diag}(\mathrm{CM}(Y, f(X))))}{\mathrm{sum}(\mathrm{CM}(Y, f(X)))}$$

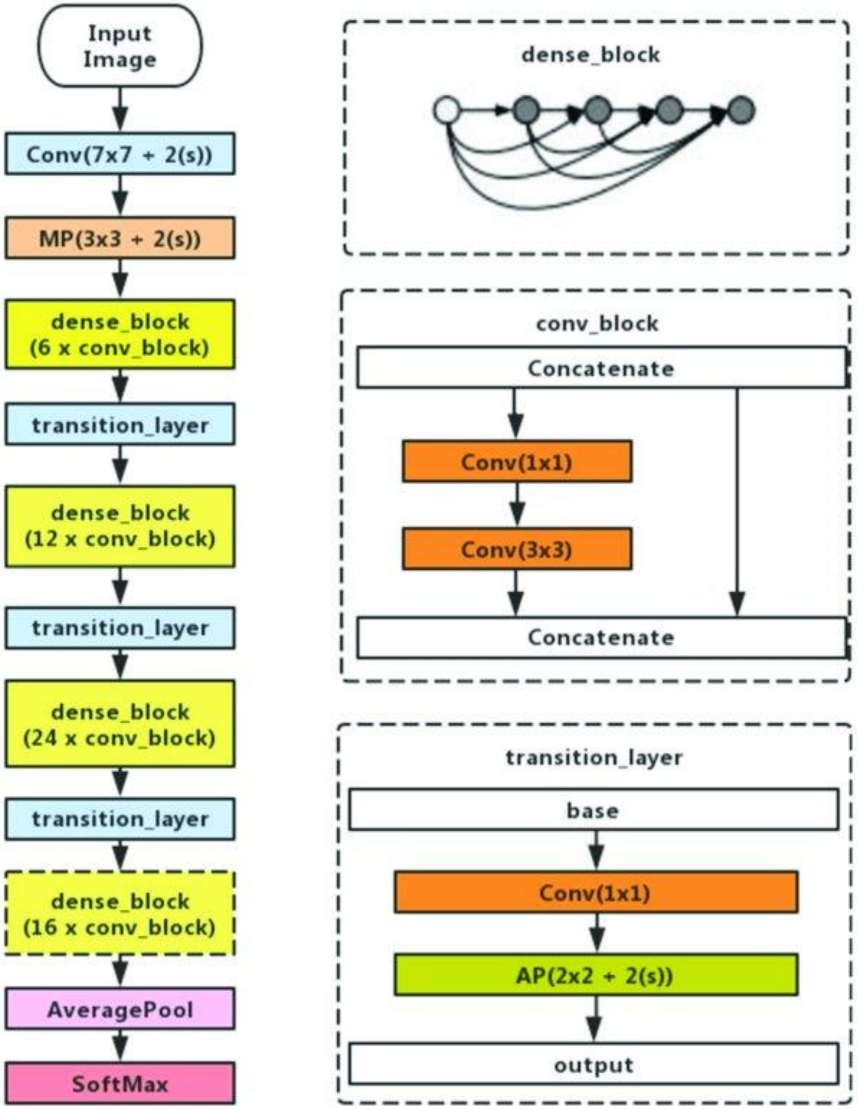
    - F1-score (landmarks)
    - Confusion matrix (between landmarks)
    - Tool: PyTorch

Deep Learning-based anatomical site classification

- DenseNet-121
  - Multi-class cross-entropy loss:

$$L(\hat{y}, y) = -\sum_{k=1}^{K} y^{(k)} \log \hat{y}^{(k)}$$

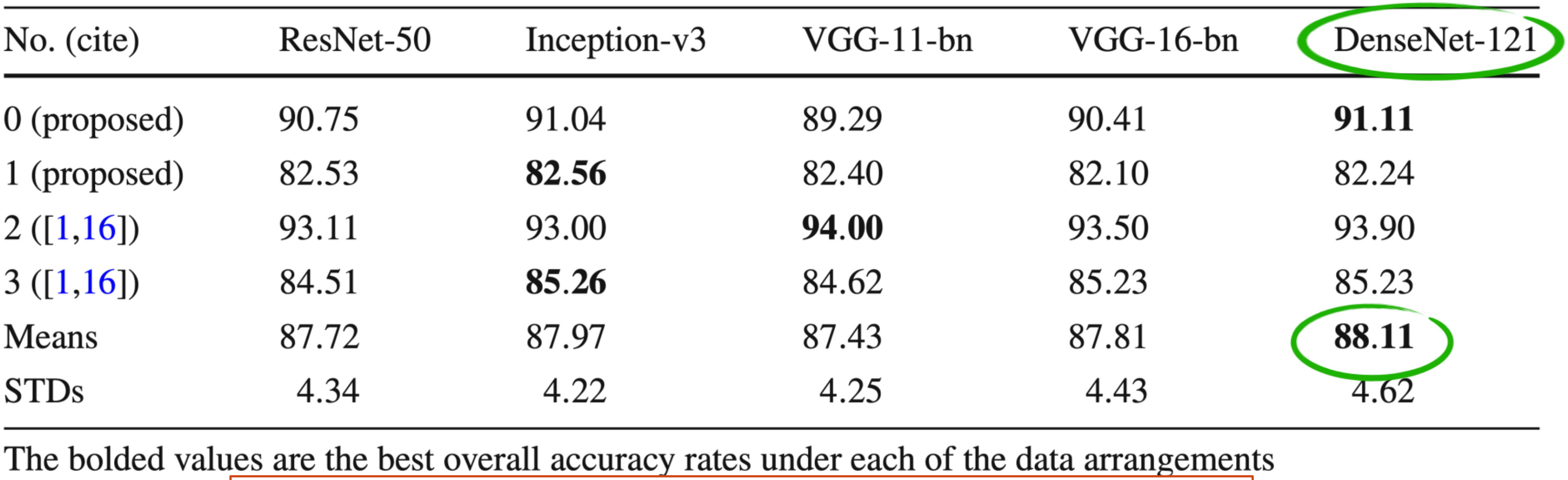
  - Data augmentation: Rotation, flipping, random value shifting, random scaling, colour jitter



[Ji et al., '19]

Results

- Evaluation of the CNN models
    - The average overall accuracy of these four models shows that DenseNet-121 gave slightly better accuracy
    - All CNN models performed equally good that demonstrate their strong learning capability and the practicality of our anatomical classification guideline



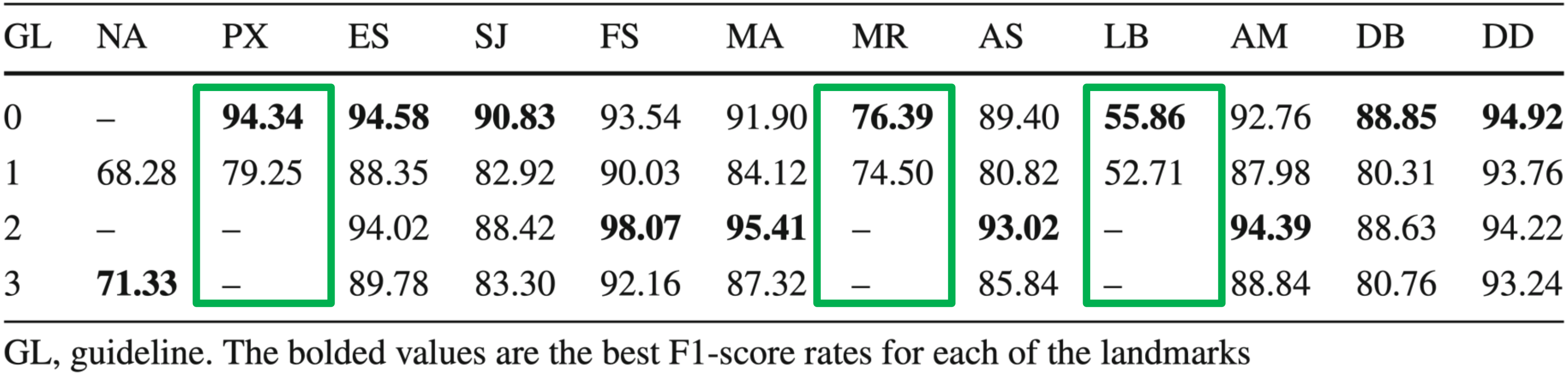
No. (cite)	ResNet-50	Inception-v3	VGG-11-bn	VGG-16-bn	DenseNet-121
0 (proposed)	90.75	91.04	89.29	90.41	**91.11**
1 (proposed)	82.53	**82.56**	82.40	82.10	82.24
2 ([1,16])	93.11	93.00	**94.00**	93.50	93.90
3 ([1,16])	84.51	**85.26**	84.62	85.23	85.23
Means	87.72	87.97	87.43	87.81	**88.11**
STDs	4.34	4.22	4.25	4.43	4.62

The bolded values are the best overall accuracy rates under each of the data arrangements

Overall accuracy (%) of five CNN models for four datasets

Results

- Evaluation of the guideline
  - The proposed guideline helps the CNN model to recognise three additional landmarks (PX, MR and LB) than the British guideline.



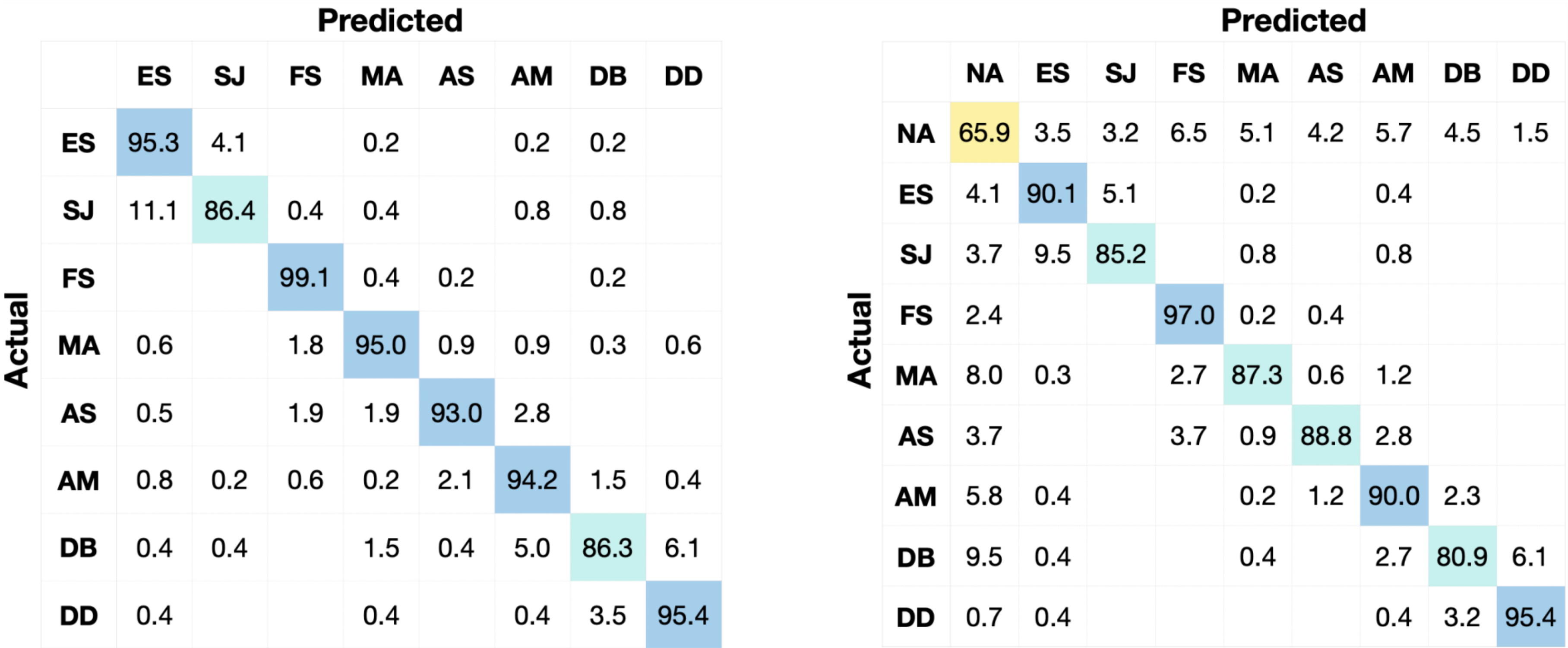
GL	NA	PX	ES	SJ	FS	MA	MR	AS	LB	AM	DB	DD
0	–	**94.34**	**94.58**	**90.83**	93.54	91.90	**76.39**	89.40	**55.86**	92.76	**88.85**	**94.92**
1	68.28	79.25	88.35	82.92	90.03	84.12	74.50	80.82	52.71	87.98	80.31	93.76
2	–	–	94.02	88.42	**98.07**	**95.41**	–	**93.02**	–	**94.39**	88.63	94.22
3	**71.33**	–	89.78	83.30	92.16	87.32	–	85.84	–	88.84	80.76	93.24

GL, guideline. The bolded values are the best F1-score rates for each of the landmarks

The F1-score (%) of DenseNet-121 on four datasets

Results

- Evaluation of the guideline
  - The CNN model evaluated on our trimmed dataset corresponding to the British guideline (since NA, PX, MR and LB are excluded) achieved superior performance



Predicted

Actual

	ES	SJ	FS	MA	AS	AM	DB	DD
ES	95.3	4.1		0.2		0.2	0.2	
SJ	11.1	86.4	0.4	0.4		0.8	0.8	
FS			99.1	0.4	0.2		0.2	
MA	0.6		1.8	95.0	0.9	0.9	0.3	0.6
AS	0.5		1.9	1.9	93.0	2.8		
AM	0.8	0.2	0.6	0.2	2.1	94.2	1.5	0.4
DB	0.4	0.4		1.5	0.4	5.0	86.3	6.1
DD	0.4			0.4		0.4	3.5	95.4

Predicted

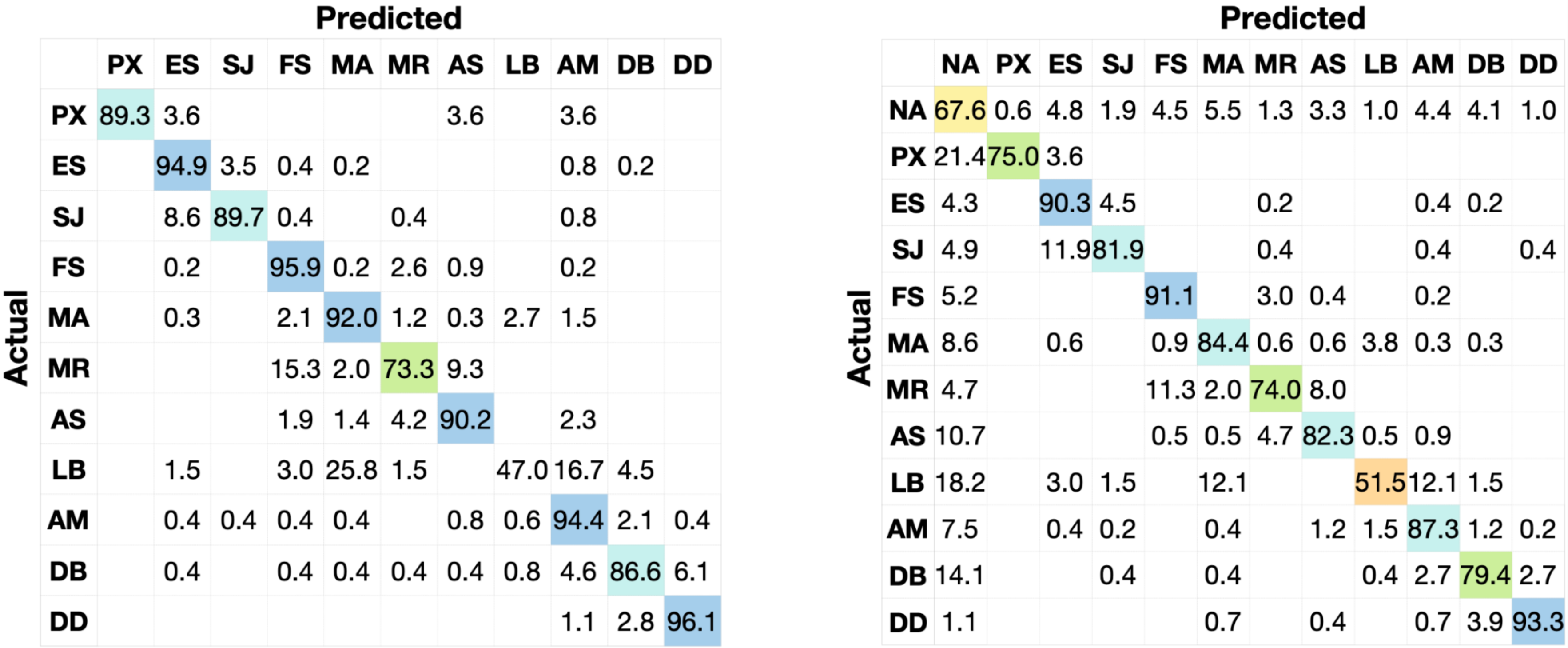
Actual

	NA	ES	SJ	FS	MA	AS	AM	DB	DD
NA	65.9	3.5	3.2	6.5	5.1	4.2	5.7	4.5	1.5
ES	4.1	90.1	5.1		0.2		0.4		
SJ	3.7	9.5	85.2		0.8		0.8		
FS	2.4			97.0	0.2	0.4			
MA	8.0	0.3		2.7	87.3	0.6	1.2		
AS	3.7			3.7	0.9	88.8	2.8		
AM	5.8	0.4			0.2	1.2	90.0	2.3	
DB	9.5	0.4			0.4		2.7	80.9	6.1
DD	0.7	0.4					0.4	3.2	95.4

Confusion matrix for the model based on the British guideline

Results

- Evaluation of the guideline
  - The performance is low for LB (class 7) since it is hard to find a reference to well recognise LB from a single image



Predicted

Actual

	PX	ES	SJ	FS	MA	MR	AS	LB	AM	DB	DD
PX	89.3	3.6					3.6		3.6		
ES		94.9	3.5	0.4	0.2				0.8	0.2	
SJ		8.6	89.7	0.4		0.4			0.8		
FS		0.2		95.9	0.2	2.6	0.9		0.2		
MA		0.3		2.1	92.0	1.2	0.3	2.7	1.5		
MR				15.3	2.0	73.3	9.3				
AS				1.9	1.4	4.2	90.2		2.3		
LB		1.5		3.0	25.8	1.5		47.0	16.7	4.5	
AM		0.4	0.4	0.4	0.4		0.8	0.6	94.4	2.1	0.4
DB		0.4		0.4	0.4	0.4	0.4	0.8	4.6	86.6	6.1
DD									1.1	2.8	96.1

Predicted

Actual

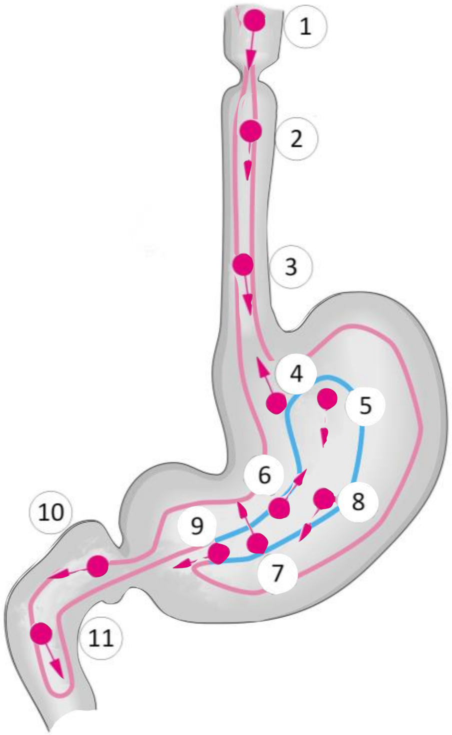
	NA	PX	ES	SJ	FS	MA	MR	AS	LB	AM	DB	DD
NA	67.6	0.6	4.8	1.9	4.5	5.5	1.3	3.3	1.0	4.4	4.1	1.0
PX	21.4	75.0	3.6									
ES	4.3		90.3	4.5			0.2			0.4	0.2	
SJ	4.9		11.9	81.9			0.4			0.4		0.4
FS	5.2				91.1		3.0	0.4		0.2		
MA	8.6		0.6		0.9	84.4	0.6	0.6	3.8	0.3	0.3	
MR	4.7				11.3	2.0	74.0	8.0				
AS	10.7				0.5	0.5	4.7	82.3	0.5	0.9		
LB	18.2		3.0	1.5		12.1			51.5	12.1	1.5	
AM	7.5		0.4	0.2		0.4		1.2	1.5	87.3	1.2	0.2
DB	14.1			0.4		0.4			0.4	2.7	79.4	2.7
DD	1.1					0.7		0.4		0.7	3.9	93.3

Confusion matrix for the model based on proposed guideline

Discussion

➢ Successful points

- Small amount of data required for training model
- Annotation friendly
- Adapted to the British examination procedure
- Recognize 3 more landmarks that the British guideline
- Enable photo-documentation of upper GI endoscopy



Discussion

Issues

- We observe the errors from the confusion matrices
  - Cause:
    - No temporal information
    - Several landmarks with similar tissue appearances are easily misclassified to each other
  - Solution:
    - To further improve the results, we plan to analyse EGD videos in future using 3D CNN and recurrent neural networks, which will incorporate both spatial feature representation and temporal information simultaneously

Discussion

Issues

- Class NA was confused with the other landmarks
  - Cause:
    - NA and the other landmarks shared several features
    - There is no clear boundary between blurry landmarks and NA
  - Solution:
    - Train a special classifier to divide the NA and the others into two classes. And then train another classifier to recognize each useful landmark.

Conclusion

- A modified guideline for upper GI endoscopic photo-documentation
- A new upper GI endoscopic dataset
- A complete workflow for EGD image classification

Thank you very much for your attention