CVPR
JUNE 17-21, 2024
SEATTLE, WA

FocusMAE: Gallbladder Cancer Detection from Ultrasound Videos with Focused Masked Autoencoders





Soumen Basu [1*]   Mayuna Gupta [1*]   Chetan Madan [1]   Pankaj Gupta [2]   Chetan Arora [1]

[1] Indian Institute of Technology, Delhi
[2] Postgraduate Institute of Medical Education & Research, Chandigarh
* Joint first-authors

Gallbladder Cancer(GBC)

- Nearly 85,000 deaths every year worldwide
- 5 year survival rate is 5%
- Mean survival – 6 months (patients at advanced stage)
- Quick Metastasis due to adjacent contiguous liver tissues
- Silent progress – often detected at a very late stage
- Early detection and timely surgery – to improve the survival statistics

Why Ultrasound (US) Videos?

- US is the most common imaging modality for abdominal ailments – scans are collected as video
- Highly accessible and low cost – excellent candidate modality for GBC detection
- No existing work on AI-based GBC detection from US Videos prior to our work
- Previous works are based on Image based techniques, requiring radiologists to select the key informative frames from a US video – observer bias, additional work
- Single frames may lack sufficient information for capturing disease manifestation

Major Challenges

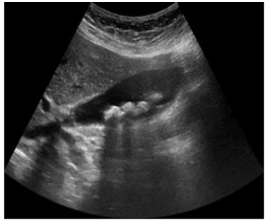
- Anatomy
  - Non-regular anatomy of malignant gallbladder (loss of interface with adjacent organs, irregular anatomical structure)
- Low Image Quality
  - Noise, artifacts such as shadow, and spurious textures
- Hand-held Sensor – Observer Bias
  - High degree of variability across radiologists, and medical centers
- Visual features of GBC can be similar to benign conditions

Challenges

Anatomy

Normal GB

Benign GB

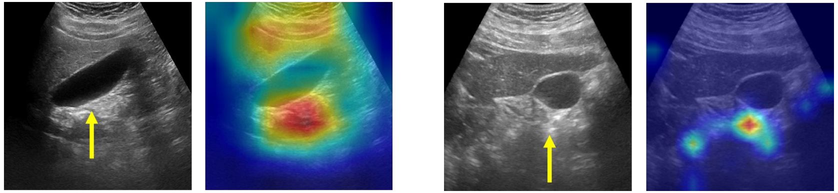


Malignant GB

Normal, Benign GB - regular anatomy

Malignant - clear boundary is absent

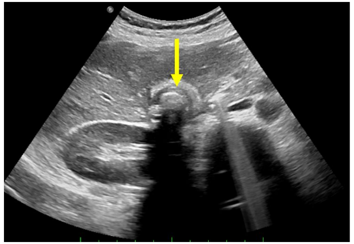
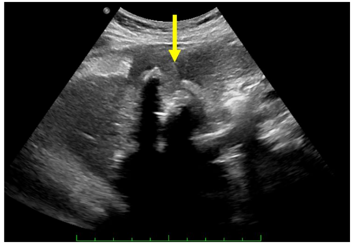
Texture and Noise



DNNs often get biased by the adjacent organ tissues and spurious echogenic textures – focuses on textures instead of the GB

Challenge

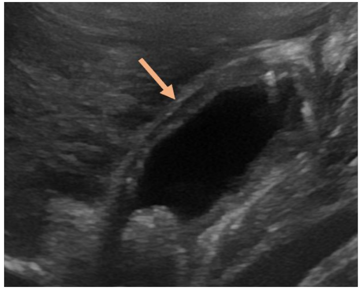
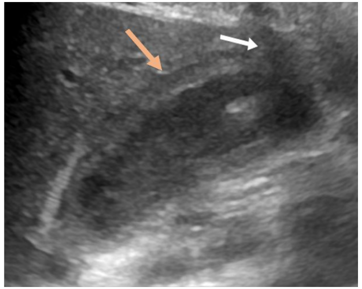
Observer Variance

Due to the handheld sensor, the scanning plane may change – introduce observer bias

Left and right shows same GB from different scanning views – drastic change

Confounding Clinical Characterizations

Benign (left) GB wall thickening usually presents layered appearance.

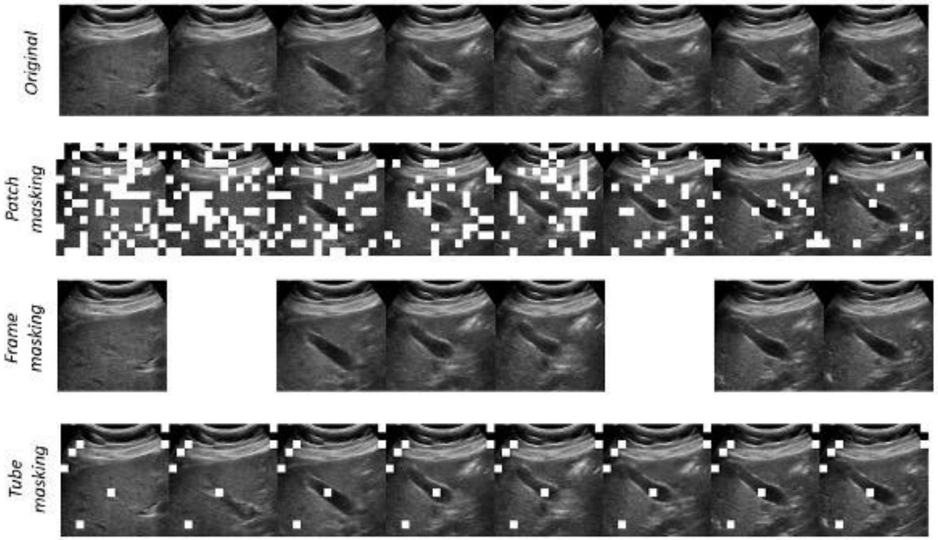
Malignant (right) GB wall thickening can sometimes show such layered appearances

MAE Recap

- Masked Autoencoders (MAEs) mask certain parts of the input and try to reconstruct it
  - Minimize reconstruction loss
  - Learn representation

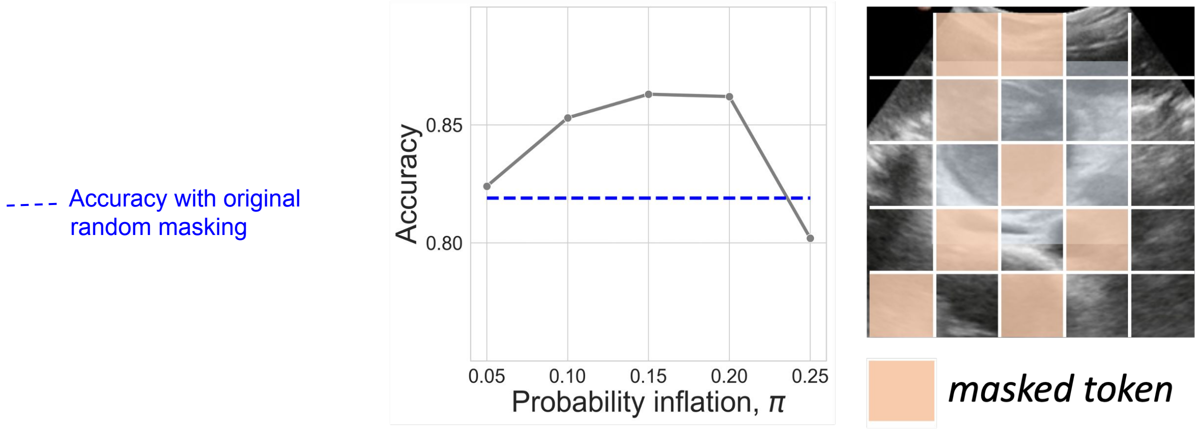
- Most SOTA MAEs use random masking in images or videos

- Random masking is not robust to small pathology areas with large background (low info) regions
  - May end up learning the background representation



Various common random masking strategies

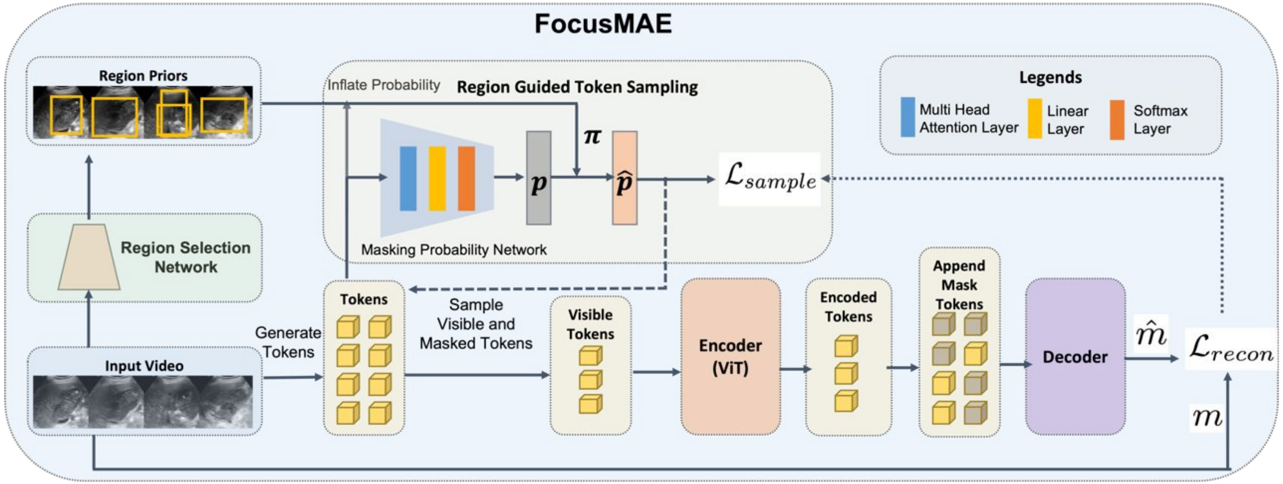
Idea: Selectively Bias Masking Probability



- Inflate the masking probabilities of the Regions of Interest (ROI) by π – adaptively mask and reconstruct high information ROI – robust representation learning

- Excessive masking of ROI degrades performance – use learnable sampling probabilities

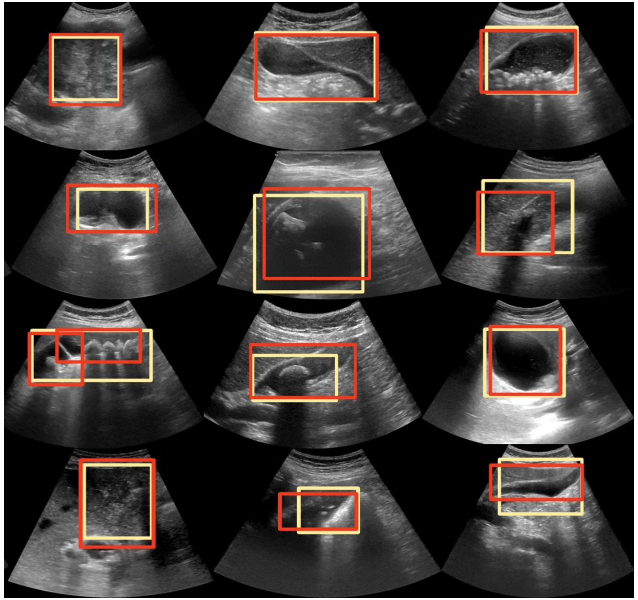
Our Solution: FocusMAE



- Use Object Detector to generate high information region priors (candidate ROI)
- Bias the masking probability of the tokens within ROI to learn representation of the pathology/ disease

Region Selection Network

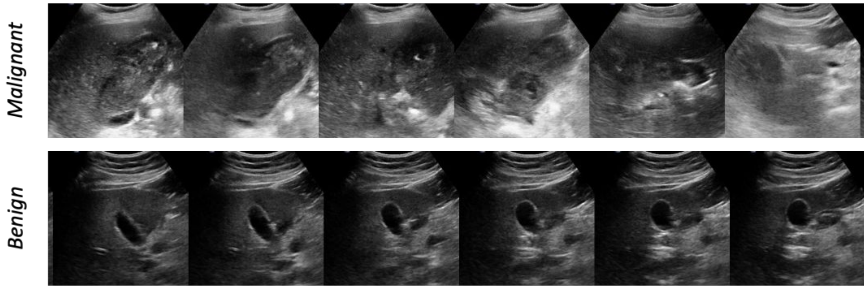
Model	mIoU	Precision	Recall
Faster-RCNN	71.1 ± 2.7	96.0 ± 2.6	99.2 ± 0.7
YOLOv4	70.7 ± 2.9	98.1 ± 2.3	97.9 ± 1.5
CentripetalNet	60.4 ± 4.7	95.1 ± 3.8	89.6 ± 7.3
Reppoints	69.1 ± 3.2	95.2 ± 3.9	99.7 ± 0.4



- Detectors only select GB vs background - ROI
- Faster-RCNN achieves best mIoU with very high recall and precision

Dataset

- We contribute 27 malignant video samples to the publicly available GBUSV [1] dataset
- Dataset 91 videos
  - 59 malignant (41 patients) and 32 benign (32 patients)
  - 5-fold cross-validation (patient-level splits)

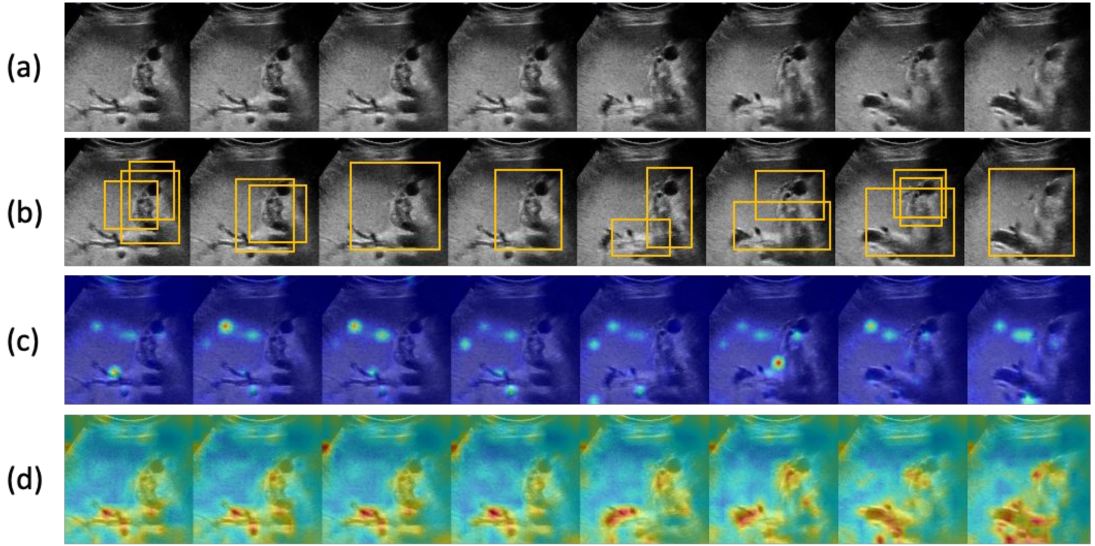


[1] Basu et al. "Unsupervised Contrastive Learning of Image Representations from Ultrasound Videos with Hard Negative Mining" MICCAI 2022.

Key Results

Group	Method	Backbone	Acc.	Spec.	Sens.
Human Experts	Radiologist A	–	0.786±0.134	1.000±0.000	0.672±0.201
	Radiologist B	–	0.874±0.088	1.000±0.000	0.811±0.126
Image-based	ResNet50 [25]	CNN	0.711±0.091	0.822±0.102	0.672±0.147
	InceptionV3 [43]	CNN	0.734±0.089	0.953±0.072	0.647±0.107
	Faster-RCNN [41]	CNN	0.757±0.058	0.687±0.056	0.808±0.091
	EfficientDet [44]	CNN	0.789±0.084	0.761±0.099	0.828±0.061
	ViT [13]	Transformer	0.796±0.068	0.751±0.128	0.820±0.076
	DEIT [46]	Transformer	0.829±0.034	0.787±0.154	0.845±0.058
	PVTv2 [49]	Transformer	0.831±0.041	0.857±0.167	0.834±0.068
	GBCNet [5]	CNN	0.840±0.105	0.843±0.204	0.843±0.072
	US-UCL [8]	CNN	0.808±0.127	0.871±0.217	0.776±0.109
	RadFormer (SOTA) [6]	Transformer	0.840±0.105	0.776±0.162	0.877±0.088
Video-based	Video-Swin [34]	Transformer	0.925±0.053	**1.000±0.000**	0.903±0.085
	TimeSformer [9]	Transformer	0.920±0.058	0.967±0.067	0.909±0.058
	VidTr [33]	Transformer	0.924±0.038	**1.000±0.000**	0.800±0.072
	VideoMAEv2 [48]	Transformer	0.942±0.066	0.937±0.078	0.940±0.120
	AdaMAE [4]	Transformer	0.947±0.053	0.952±0.066	0.913±0.116
	FocusMAE (Ours)	Transformer	**0.964±0.047**	0.910±0.117	**1.000±0.000**

Qualitative Analysis



- FocusMAE (d) attentions are more guided to the pathology and anatomical structures as compared to VideoMAE (c) attentions.

Generality to CT-based COVID Detection

Group	Method	Acc.	Spec.	Sens.
Image-based	ResNet50 [25]	0.721	0.739	0.711
	InceptionV3 [43]	0.672	0.739	0.632
	ViT [13]	0.770	0.783	0.763
	DEIT [46]	0.770	0.696	0.816
Video-based	TimeSformer [9]	0.700	0.739	0.474
	VideoMAE [48]	0.852	**0.956**	0.789
	FocusMAE (Ours)	**0.885**	0.895	**0.869**

On the publicly available COVID-CT-MD [1] data.

[1] Afshar et al. Covid-ct-md, covid-19 computed tomography scan dataset. Scientific Data, 2021.

Thank You!

For more details (code, dataset), please visit project website

- https://gbc-iitd.github.io/focusmae

Interested to know about the Computer Vision Group at IIT Delhi?

- Please visit: https://vision-iitd.github.io