FocusMAE: Gallbladder Cancer Detection from Ultrasound Videos with Focused Masked Autoencoders
(Supplementary Material)

Soumen Basu[1], Mayuna Gupta[1], Chetan Madan[1], Pankaj Gupta[2], Chetan Arora[1]
[1] IIT, Delhi, India [2] PGIMER, Chandigarh, India
https://gbc-iitd.github.io/focusmae

A. Region Selection Network

We have experimented with multiple deep detectors to localize the candidate regions. We haven't used the bounding box annotations in the video for training the candidate region networks. Instead, we used the public GBCU dataset to pretrain the detectors for localizing the malignancy. We then lowered the threshold to generate multiple candidate regions for the video frames used in the FocusMAE experiments.

To calculate precision and recall in the GB localization phase, following the recommendation of [6], we determine a predicted region as a true positive if its center falls within the bounding box of the ground truth region. Conversely, if the center is outside the bounding box, we categorize the prediction as a false positive attributed to localization error. Tab. S1 shows the mIoU and the recall for the different candidate region detectors.

Fig. S1 shows sample object region localization of the RPN. We adopted a FasterRCNN-based RPN for generating the candidate regions for using as priors in FocusMAE as the detector achieves the best recall rate.

Model	mIoU	Precision	Recall
Faster-RCNN	0.712	0.952	0.994
YOLO	0.767	0.979	0.962
CentripetalNet	0.614	0.947	0.909
Reppoints	0.682	0.942	0.997
DETR	0.724	0.962	0.988

Table S1. Comparison of the candidate region selection models.

B. Region Selection Network Implementation

We adopted the Faster-RCNN [5] model for candidate region selection. A frozen Resnet50 Feature Pyramid backbone is used. The input size was 800 × 1333 × 3. We used a SGD optimizer with LR = 0.005, momentum = 0.9, and weight decay = 0.0005. We used a batch size of 16 and trained for 60 epochs on the GBCU dataset.

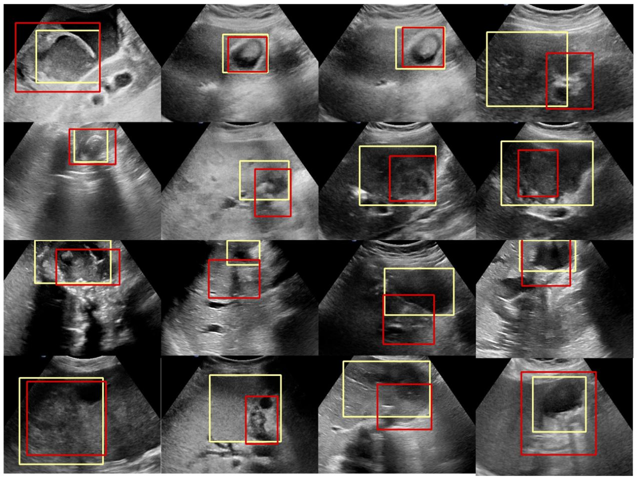

Figure S1. Sample candidate bounding boxes generated by the region selection network.

C. Visualization

Fig. S2 and Fig. S3 show the attention visuals for the proposed FocusMAE method on additional data samples. Evidently, FocusMAE is able to attend the salient regions for disease detection.

D. Baseline Implementation Details

Tab. S2 lists the configurations of all baseline models used in this study. We trained our models on 4 Nvidia Tesla V100 32GB GPUs. The table includes a brief description of the model, input sizes, optimizer parameters, other relevant hyper-parameters such as learning rate, weight decay, momentum, batch size, and the number of training epochs for the network.

For VideoMAEv2 pretraining, we used a Vision transformer (ViT) backbone, with random masked auto-

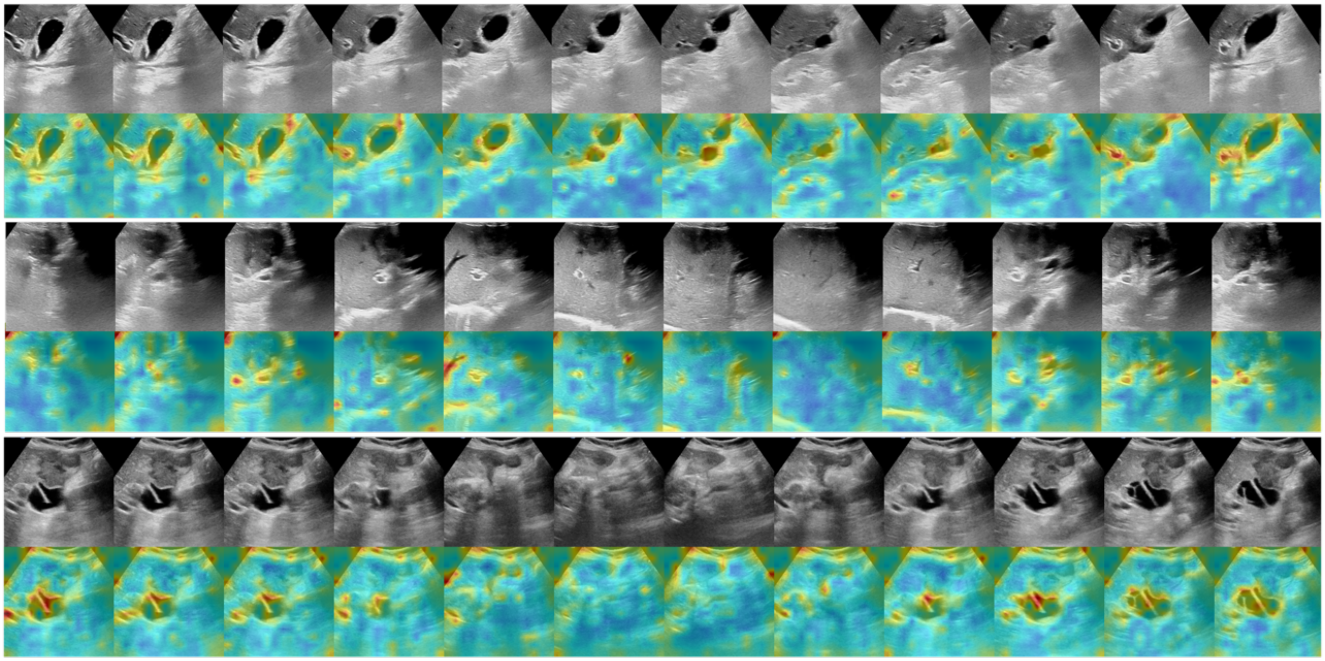

Figure S2. Attention visuals for FocusMAE for the GBC detection task on the US videos. We show three different maligant video samples. For each video sample, the upper row shows the sequence with the original frames, and the lower row shows the attention on the frames.

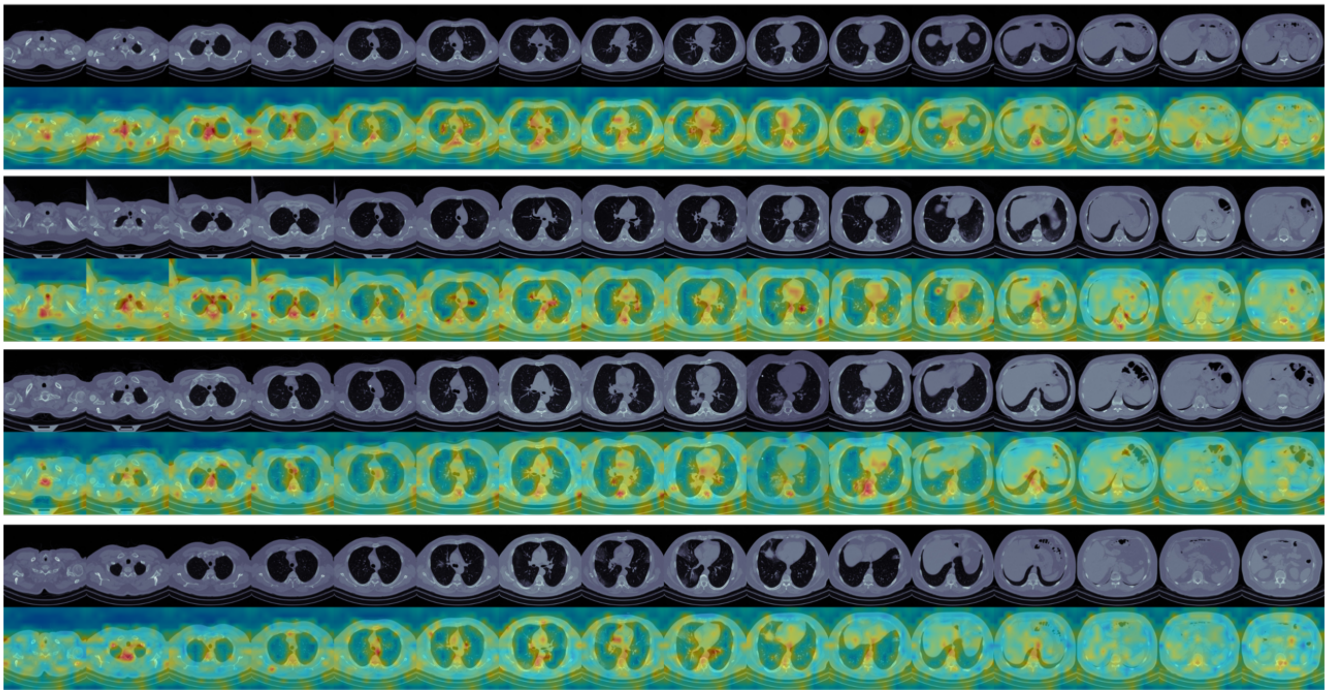

Figure S3. Attention visuals for FocusMAE for COVID detection from CT images. We show four COVID CT samples. For each sample, the upper row shows the sequence with the original CT slices, and the lower row shows the attention on these slices.

encoders. Masking was done in both encoder and decoders. All attention-based layers were trainable. We used the ViT-S model. The input size was $3 \times 16 \times 224 \times 224$. We initiated the ViT weights with the Kinetics-pretrained VideoMAE weights. We have optimized the MSE loss for original and reconstructed masked patches on the GBC US Video dataset

Model	Description	Input Size	Optimizer	Batch size	Epochs/ Steps
VideoMAEv2 [7]	Vision transformer (ViT) backbone, with random masked auto-encoders. Masking in both encoder and decoders. All attention-based layers were trainable. ViT base model used for inference.	$3 \times 16 \times 224 \times 224$	AdamW, LR = 7e-5, momentum = 0.999 ,weight decay = 0.1	4	30 epochs
TimeSformer [2]	Vision transformer based space time attention. Divided space-time attention configuration used. ViT base model used for inference.	$3 \times 8 \times 224 \times 224$	SGD, LR = 0.005, weight decay=1e-4, momentum=0.9	8	25 epochs
VideoSwin [4]	Pretrained on ImageNet-1K. SwinTransformer3D based backbone. All layers were trainable	$3 \times 8 \times 224 \times 224$	SGD, LR = 0.01, weight decay=1e-4, momentum=0.9	4	30 epochs
AdaMAE [1]	Vision transformer (ViT) backbone, with adaptively masked auto-encoders. The model embedding size provided by the authors is 768; we have pre-trained the 384 version to allow for a better fit to our data. Masking in only encoders. All attention-based layers were trainable. ViT base model used for inference	$3 \times 16 \times 224 \times 224$	AdamW, LR = 1e-6, weight decay=0.9, momentum=0.99	2	10 epochs
VidTr [3]	Transformer-based video classification with separable attention. ViT-B backbone. All layers were trainable.	$3 \times 16 \times 224 \times 224$	SGD, LR = 3e-4, weight decay=1e-5, momentum=0.9	2	40 epochs

Table S2. Implementation details for the different video-based baseline networks used for US video-based classification of Gallbladder Cancer. All details are for finetuning on the GB US video dataset. Pretraining details for VideoMAE and AdaMAE are already discussed.

using an AdamW optimizer with LR = 1e-4 and momentum = 0.95. We used a batch size of 32 and trained for 1200 epochs.

AdaMAE pretraining was an adoption of the VideoMAE pretraining procedure. We used the ViT-S backbone, with adaptively masked auto-encoders. We have pre-trained the model with embedding dimension 384 to allow for a better fit to our data. We have used masking in only encoders. All attention-based layers were trainable. Similar to VideoMAE, we initialized the weights with the Kinetics preained AdaMAE weights. We used the MSE loss and used an AdamW optimizer with LR = 1e-4 and momentum = 0.95. The input sizes are $3 \times 16 \times 224 \times 224$. We used batch size of 8 and pretrained for 500 epochs.

E. Clip-level Statistics

We have a total of 484 clips sub-sampled from the 91 videos at the fine-tuning stage. Out of these, 320 clips were from the malignant videos, and contain the malignant label as per the positive biopsy reports. All clips of a malignant video is given the malignant label. Radiologists identified 199 clips out of these 320 to be malignant. At a frame-level, radiologists identified 3212 frames exhibiting signs of malignancy.

References

[1] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14507–14517, 2023. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[3] Xinyu Li, Yanyi Zhang, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. *arXiv e-prints*, pages arXiv–2104, 2021. 3

[4] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 1

[6] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammo-

grams with deep learning. *Scientific reports*, 8(1):1–7, 2018. 1

[7] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3