Supplementary Materials for Training Auxiliary Prototypical Classifiers for Explainable Anomaly Detection in Medical Image Segmentation

In this supplementary document, we complement our paper with details of the models used in our experiments and additional qualitative results. Firstly, Section A describe in-depth details of our compared models. In Section B, we then provide additional score map visualization results.

A. Compared Models and Details

MC Dropout. Following the implementation of Bayesian SegNet[1], we applied 6 Dropout layers, which activate at both training and test time, to the conventional U-Net architecture. In the Dropout layers, we set the dropout probability of dropping a connection as 0.5. Specifically, we present the modified architecture in Table 1. For inference, we took 10 FCN test runs and averaged the prediction probabilities. Then, as we mentioned in the main paper, the entropy-based OOD score can be computed via the calibrated probabilities.

Ensemble. For the Ensemble method, we first trained 10 vanilla U-Net models for each experiment. As we computed the entropy-based OOD score in MC dropout, the prediction probabilities were averaged over the 10 models at test time.

FCDD. To employ the principle of FCDD in a fully convolutional network f, we used an auxiliary network (projection head) to convert an intermediate latent feature $f_{\text{int}}(\mathbf{x}) \in \mathbb{R}^{h \times w \times c}$ to $e(\mathbf{x}) \in \mathbb{R}^{h \times w}$. Then, we trained the network f by using a loss function $\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{fcdd}}$, where $\mathcal{L}_{\text{fcdd}}$ was formulated by following Eq. (1) in Section 2.1.

Transformation methods. As we mentioned in our main paper, we used the following operations to generate auxiliary outliers by transforming images drawn from the InD training set: 180° rotation, vertical flip, permutation, and random operations of $\{\texttt{Swap}, \texttt{Gamma}, \texttt{Deformation}\}$ in `TorchIO`. For each data sample, one of the 6 operations were applied uniformly at random, where the parameters of `TorchIO` operations were randomly chosen.

In the following, we describe the range of parameters for each `TorchIO` transform operations.

[1]Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In British Machine Vision Conference, 2019.

Layer	Architecture
ENC1	ConvBlock(I1, O64)
ENC2	Maxpool(K2, S2), ConvBlock(I64, O128)
ENC3	Maxpool(K2, S2), ConvBlock(I128, O256)
ENC4	Maxpool(K2, S2), Dropout, ConvBlock(I256, O512)
ENC5	Maxpool(K2, S2), Dropout, ConvBlock(I512, O1024), Dropout
DEC1	UpConvBlock(I1024, O512) Concat w/ ENC4 ConvBlock(I1024, O512), Dropout
DEC2	UpConvBlock(I512, O256) Concat w/ ENC3 ConvBlock(I512, O256), Dropout
DEC3	UpConvBlock(I256, O128) Concat w/ ENC2 ConvBlock(I256, O128), Dropout
DEC4	UpConvBlock(I128, O64) Concat w/ ENC1 ConvBlock(I128, O64)
DEC5	Conv(I64, O4, K1, S1, P0)

Table 1. A modified U-Net architecture for MC Dropout. We added 6 Dropout layers to the vanilla U-Net model. In each component, I and O denote the number of input and output channels, where K, P, and S indicate kernel, padding, and stride sizes, respectively.

- Deformation: `max_displacement` $= 40$ and `num_control_points` $\in \{10, 12, 14, 16, 18\}$
- Gamma: `log_gamma` $\in \{0.8, 1.0, 1.2, 1.4, 1.6\}$
- Swap: `patch_size` $\in \{30, 40, 50, 60, 70\}$ and `num_iterations` $\in \{2, 3, 4, 5, 6\}$

B. Additional Visualization of Score Map

This section provides additional visualization of images $\mathbf{x}$ and the corresponding scaled score maps $A_\xi(\mathbf{x})$ given by our model. Figures 1, 2, and 3 demonstrate the results of the models trained with the M&Ms A, M&Ms B, and PROSTATEx (Trans.) datasets, respectively. In the figures, $A_\xi(\mathbf{x})$ is overlaid on $\mathbf{x}$ via bilinear interpolation, where blue and red colors correspond to 0 and 1 values in $A_\xi(\mathbf{x})$, respectively. In other words, the red-colored regions can be anomalous (having potential errors) in the perspective of our FCN f.

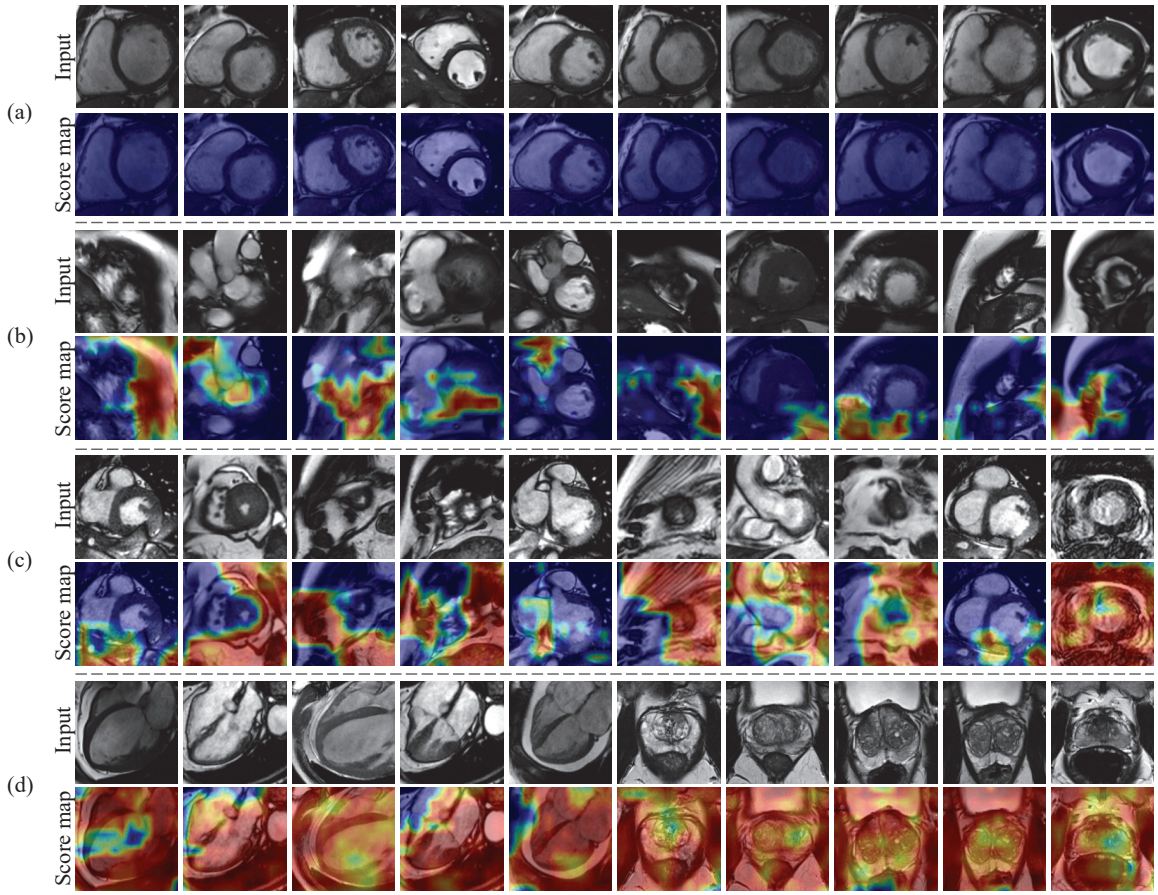

Figure 1. A visualization of images $\mathbf{x}$ and their scaled score maps $A_\xi(\mathbf{x})$, where the model is trained with the M&Ms A dataset. We depict (a) in-distribution test images having no anomalous regions, (b) in-distribution test images having anomalous sub-regions, (c) OoD images of C3 (M&Ms B and M&Ms C/D), and (d) OoD images of C1 and C2 (M&Ms-2 LA and PROSTATEx (Trans.)).

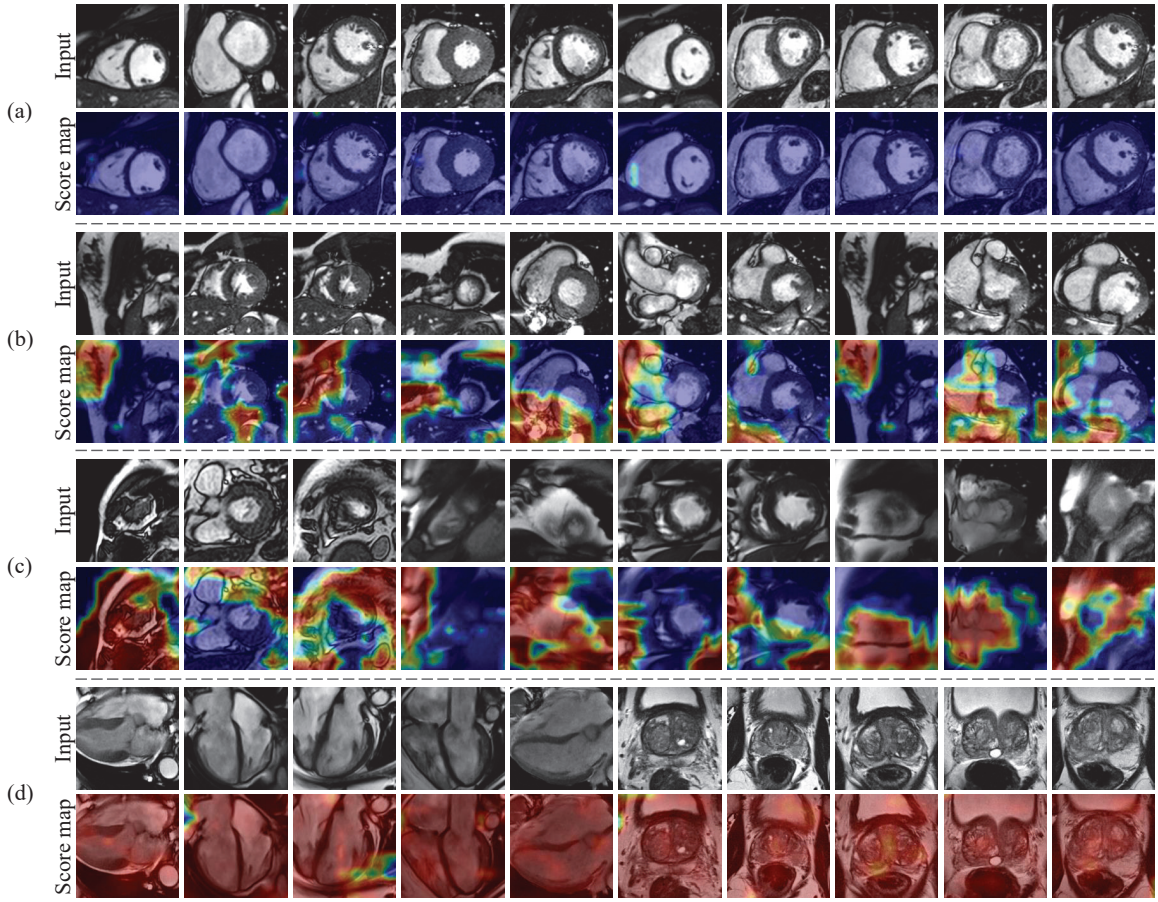

Figure 2. A visualization of images $\mathbf{x}$ and their scaled score maps $A_\xi(\mathbf{x})$, where the model is trained with the M&Ms B dataset. We depict (a) in-distribution test images having no anomalous regions, (b) in-distribution test images having anomalous sub-regions, (c) OoD images of C3 (M&Ms A and M&Ms C/D), and (d) OoD images of C1 and C2 (M&Ms-2 LA and PROSTATEx (Trans.)).

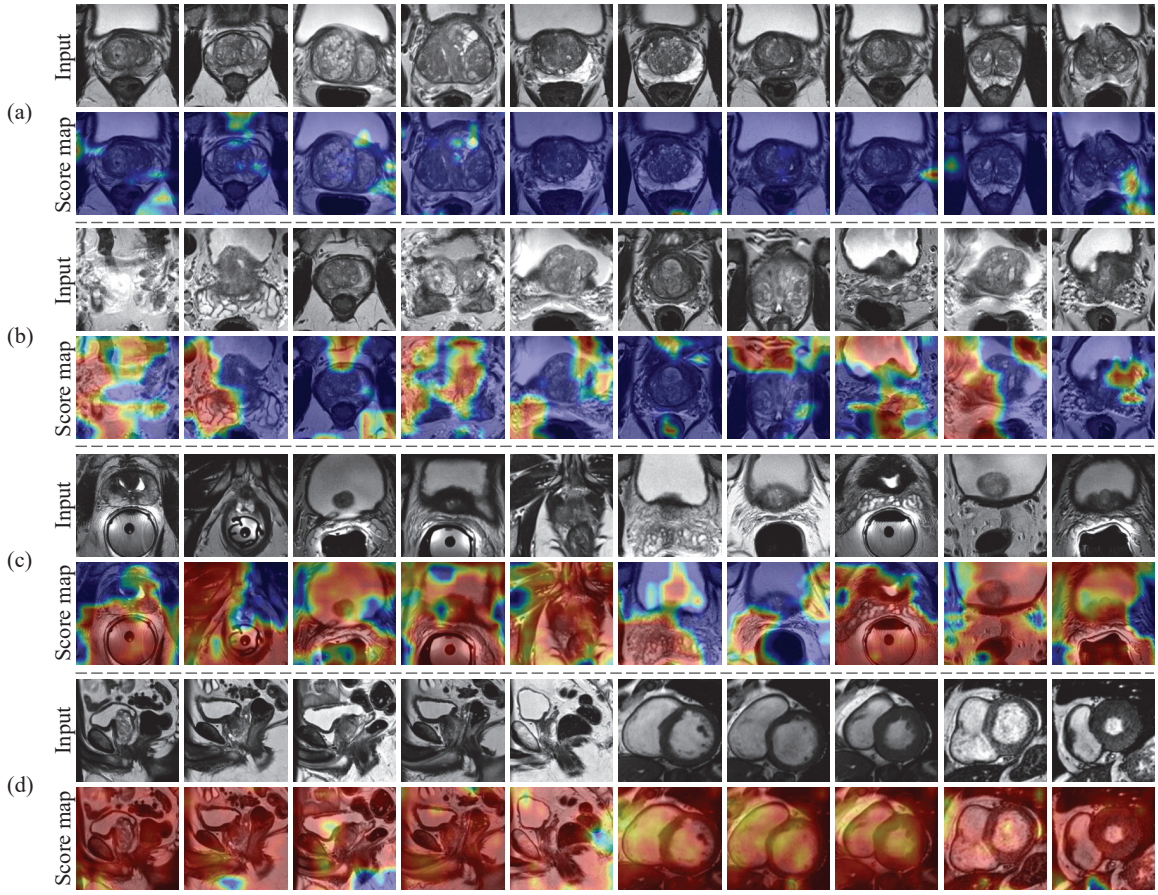

Figure 3. A visualization of images $\mathbf{x}$ and their scaled score maps $A_\xi(\mathbf{x})$, where the model is trained with the PROSTATEx (Trans.) dataset. We depict (a) in-distribution test images having no anomalous regions, (b) in-distribution test images having anomalous sub-regions, (c) OoD images of C3 (PROMISE12), and (d) OoD images of C1 and C2 (PROSTATEx (Sag.), M&Ms A and M&Ms B).