Supplementary Material for XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations

Eunji Kim[1] Siwon Kim[1] Minji Seo[1] Sungroh Yoon[1,2,]
[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2] ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University
{kce407, tuslkkk, minjiseo, sryoon}@snu.ac.kr

A. Training Details

We employ the Adam [3] optimizer with a learning rate of 10^{-4} for the pretrained convolutional layers and the classification layer, and 10^{-3} for the feature and occurrence modules in the feature extractor and prototype layer. The learning rates for the feature extractor and prototype layer are reduced by 0.1 after the fourth step.

The proposed method is implemented using PyTorch [4] on NVIDIA TESLA V100 and Quadro RTX 8000 GPUs.

B. Additional Results

We present additional results to support the experimental results in the main paper.

B.1. Comparison with Baselines

We compare the predictions by XProtoNet with those by the baselines (Figure A1 and A2). In Figure A1, the hernia prototype of XProtoNet and the occurrence areas of the input X-ray images show equivalent locations within the images, and the similarity scores between them are high, diagnosing appropriately. The prototype of baseline Patch$_{3\times3}$, which shows the highest performance among the baselines, seems to present the location akin to the prototype of XProtoNet, but then it erroneously outputs a different location as being similar to that of the prototype in Figure A1(a). The hernia prototypes of the other baselines Patch$_{r\times r}$ present different areas within the images, and those baselines fail to diagnose hernia in both Figures A1(a) and (b). The baseline GAP shows a high similarity score between the input X-ray image and the prototype, but does not provide a satisfactory explanation. Note that there is no bounding box annotation for hernia in the dataset.

In Figure A2, all prototypes present the region of the heart in X-ray images. However, the prototypes of the baselines show a portion of the heart or include an irrelevant part. Since the learned prototypes of the baselines are not well-matched with the input X-ray image, the baselines fail to diagnose in Figure A2(a). By contrast, XProtoNet successfully identifies cardiomegaly and shows more interpretable explanations in both Figures A2(a) and (b).

The experimental results show that XProtoNet achieves higher diagnostic performance and yields more interpretable explanations than the baselines by learning the prototype from an adaptive area and comparing it to the appropriate area on an input X-ray image.

B.2. Explanation with Prototypes

We provide additional examples of the predictions and explanations by XProtoNet (Figures A3, A4, and A5). The prototypes are visualized with the X-ray images from training data and their corresponding occurrence maps. The occurrence maps are upsampled to the input image size and normalized with the maximum value for visualization.

Figure A3 shows the diagnosis on positive and negative samples by XProtoNet using ResNet-50 [1] as a backbone. The occurrence maps show that the predicted occurrence areas for positive and negative samples are analogous, and the similarity scores of the occurrence areas with the prototypes are high for positive samples and low for negative samples, resulting in accurate diagnostic results. Figures A4 and A5 show the additional examples on positive samples from XProtoNet using ResNet-50 [1] and DenseNet-121 [2] as backbones, respectively.

B.3. XProtoNet with Prior Condition

Figure A6 shows a comparison of the explanations by XProtoNet trained with and without the prior condition. Since the prototypes of XProtoNet trained with the prior condition learn specific features from the actual signs of the disease, the corresponding occurrence areas of the input X-ray images are predicted to be suitable to detect similar features with the prototypes. It shows that it is possible to build a more reliable diagnostic system with XProtoNet by directly affecting the global explanation, which is a crucial factor for the diagnosis, rather than individually acting on each data point.

Table A1. AUC scores of XProtoNet depending on the number of prototypes per class, K. ResNet-50 is used as a backbone.

K	1	2	3	4	5	7
Mean AUC	0.812	0.815	0.820	0.820	0.817	0.816

B.4. The number of prototypes

We analyze the sensitivity of the diagnostic performance to the number of prototypes per class, K. A large K is not necessary because the characteristics that appear as a sign of a disease are not as diverse as a general object. Rather, an increase in K can cause the training signals to disperse for the prototype, which can lead to performance degradation. The value of K should be determined considering the diversity of the disease evidence. Table A1 presents the diagnostic performance of XProtoNet, which is based on ResNet-50, at different values of K. Since the performance is best at $K = 3$ and 4, we set $K = 3$ for all experiments.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[2] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[4] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

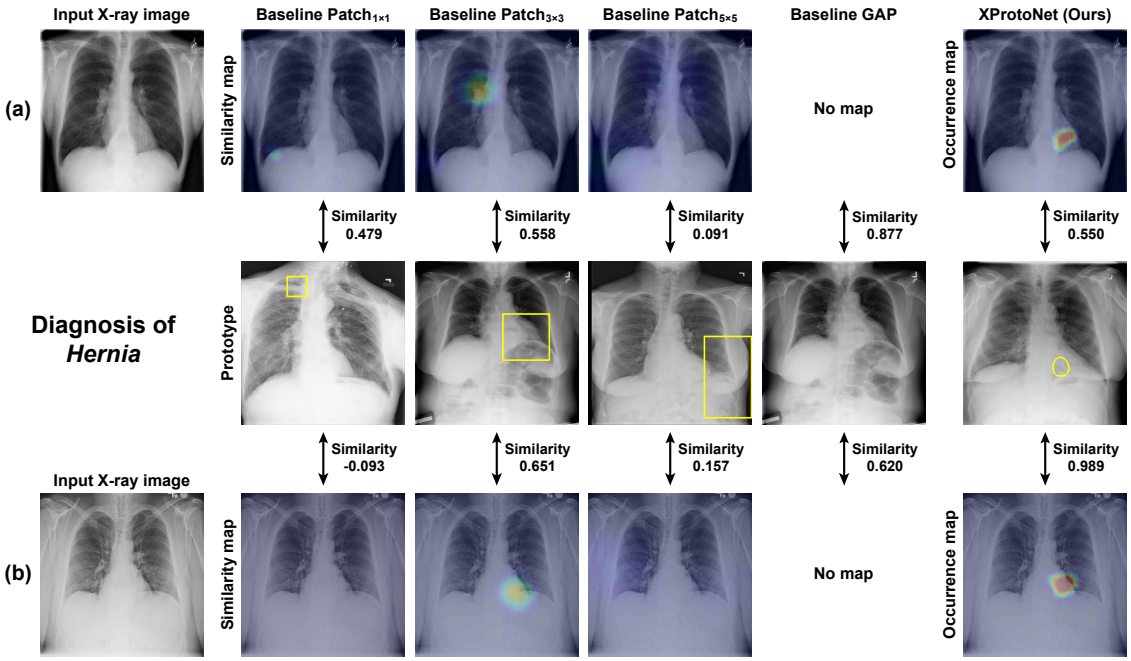

Figure A1. Comparison of the predictions by XProtoNet and the baselines for hernia diagnoses. Both input X-ray images (a) and (b) are positive samples for hernia. We show similarity maps for the baselines Patch$_{r\times r}$ and occurrence map for XProtoNet. The baseline GAP does not provide a similarity map since the features of the entire images are compared. The heatmaps are upsampled to the size of the input image. Yellow boxes and contours show the prototypes. The prototype of the baseline GAP is the entire image.

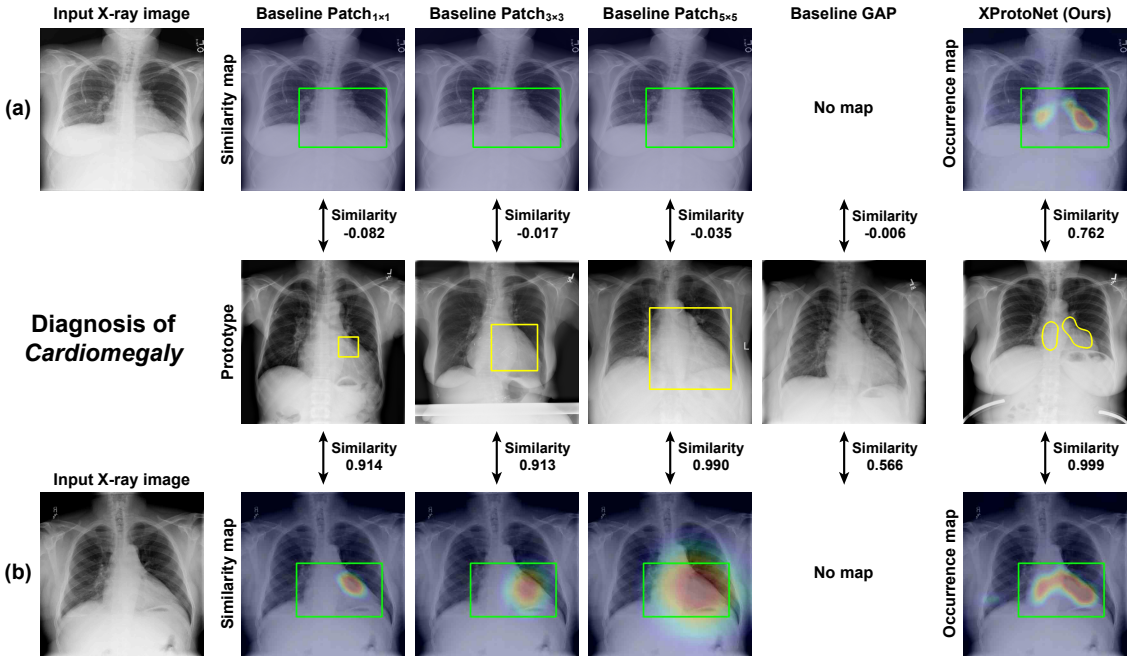

Figure A2. Comparison of the predictions by XProtoNet and the baselines for cardiomegaly diagnoses. Both input X-ray images (a) and (b) are positive samples for cardiomegaly. We show similarity maps for the baselines Patch$_{r\times r}$ and occurrence map for XProtoNet. The baseline GAP does not provide a similarity map since the features of the entire images are compared. The heatmaps are upsampled to the size of the input image. Yellow boxes and contours show the prototypes, and the green box denotes the ground-truth bounding box from the dataset. The prototype of the baseline GAP is the entire image.

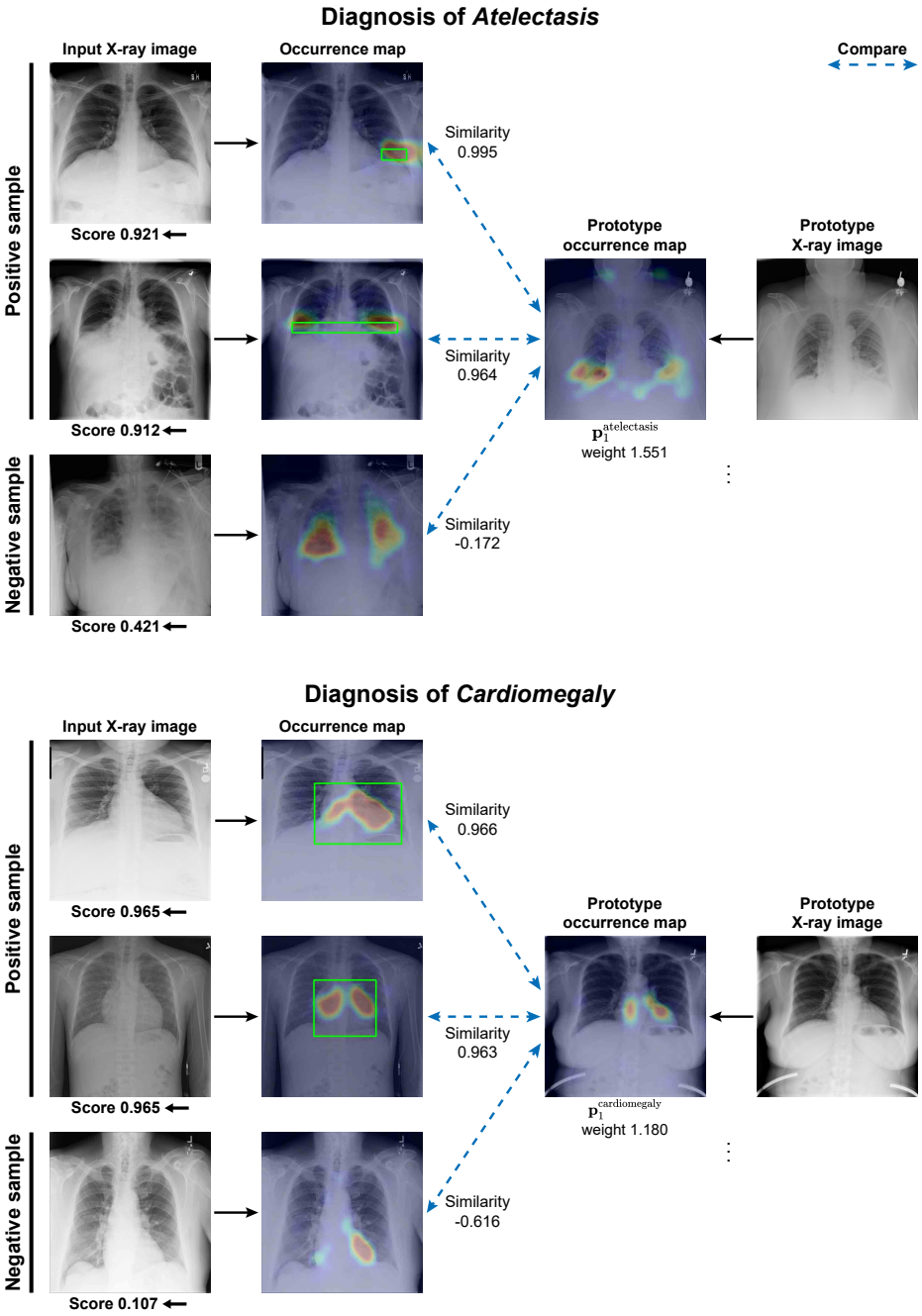

Figure A3. Examples of predictions and explanations by XProtoNet based on ResNet-50 for atelectasis and cardiomegaly diagnoses. Green boxes denote the ground-truth bounding boxes from the dataset.

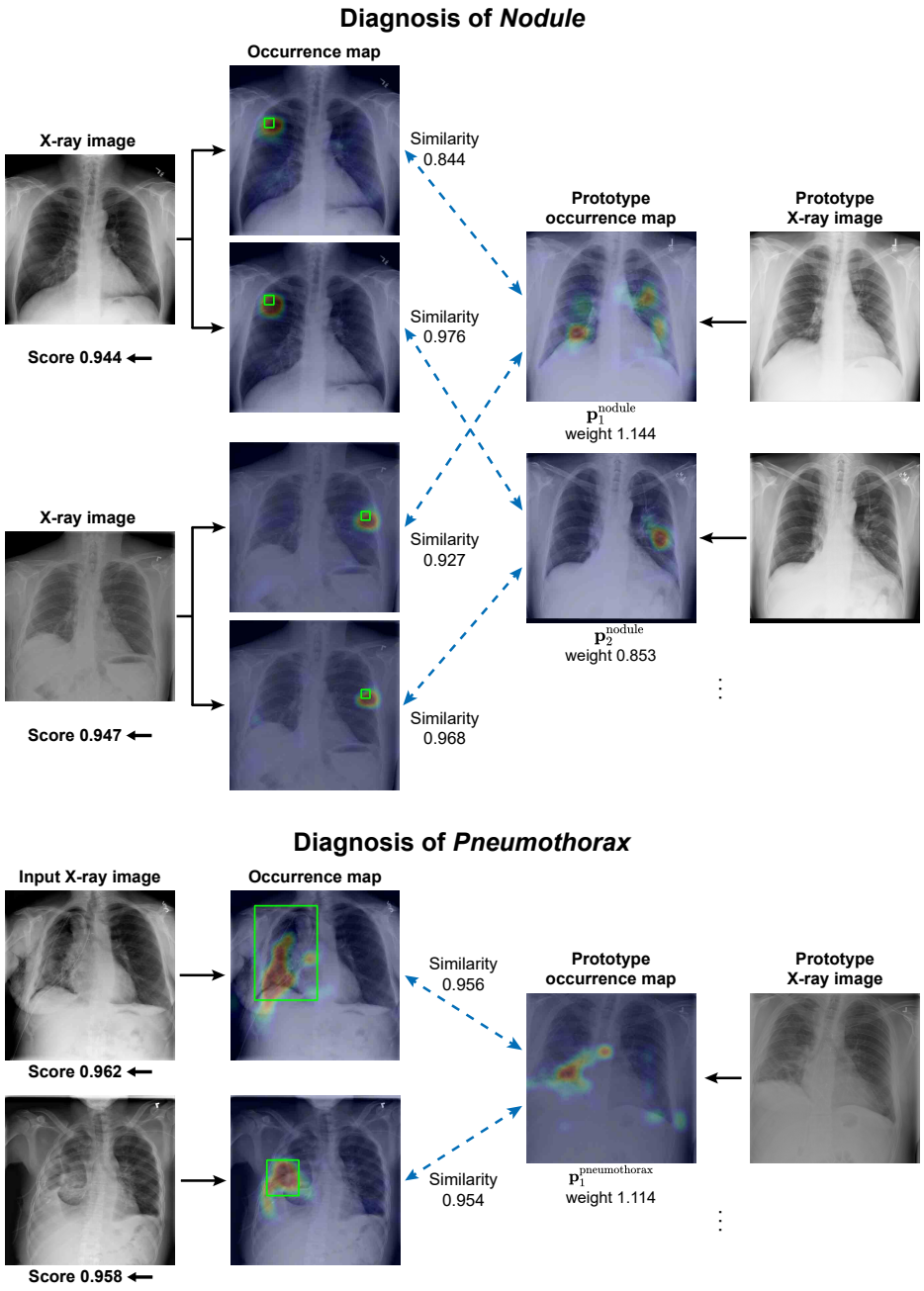

Figure A4. Examples of predictions and explanations by XProtoNet based on ResNet-50. Green boxes denote the ground-truth bounding boxes from the dataset.

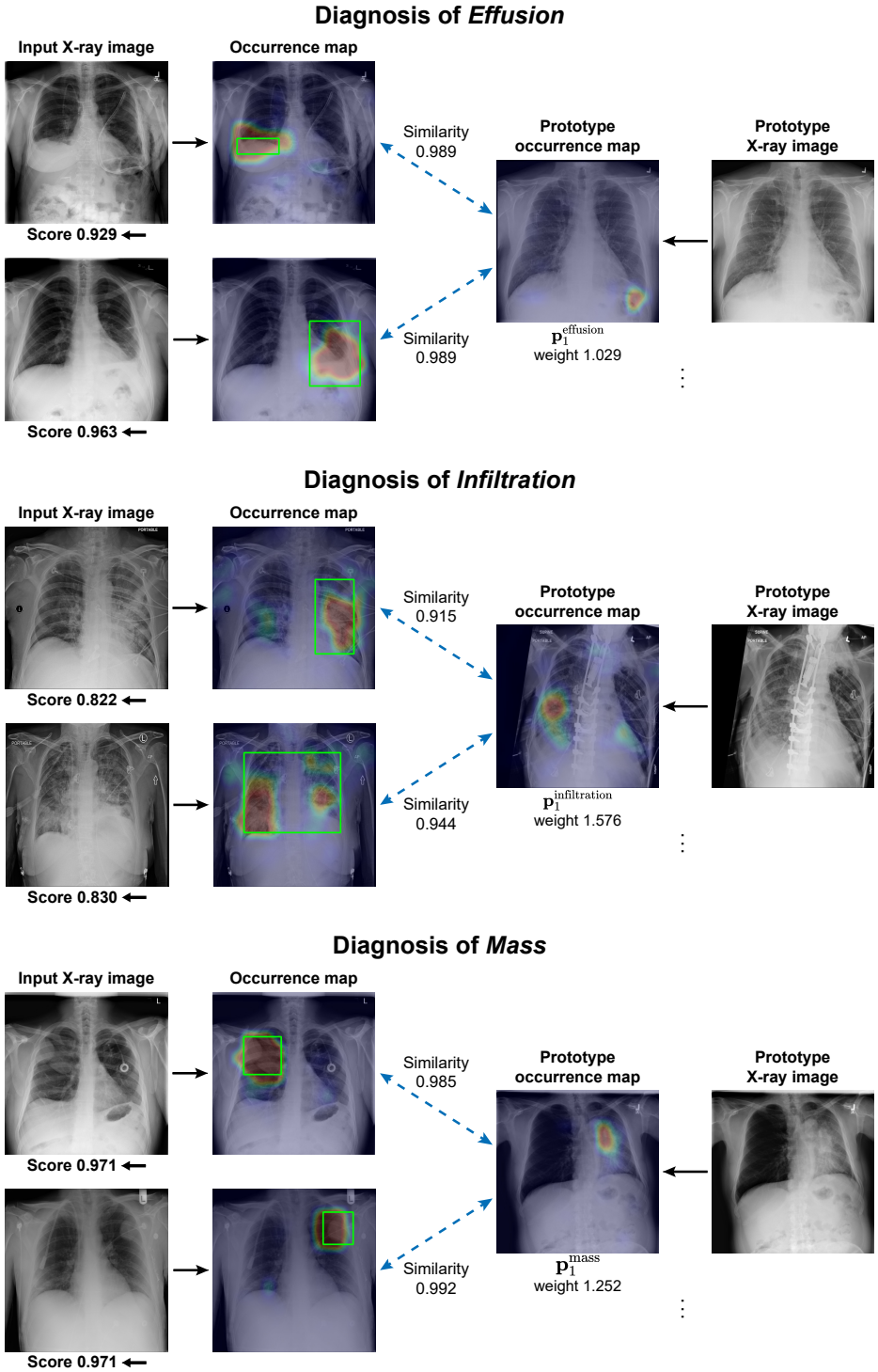

Figure A5. Examples of predictions and explanations by XProtoNet based on DenseNet-121. Green boxes denote the ground-truth bounding boxes from the dataset.

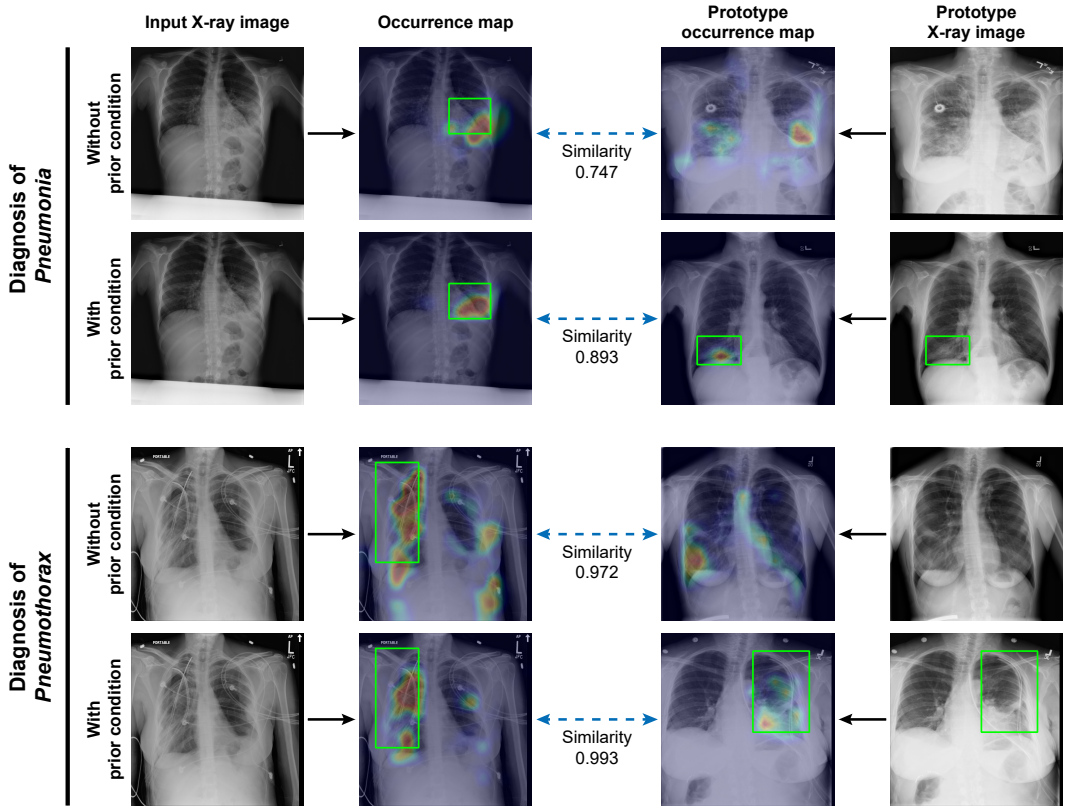

Figure A6. Examples of explanations by XProtoNet trained with and without the prior condition. Green boxes denote the ground-truth bounding boxes from the dataset.