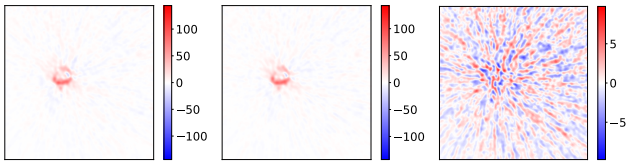
A. Appendix

Contents

A.1. About $\hat{\mathbf{u}}(\hat{\mathbf{n}})$ and $\hat{\mathbf{v}}(\hat{\mathbf{n}})$

$\mathcal{R}_\Sigma$, $\mathcal{R}_{g_\theta}$, and thus the output of our method, are dependant on the choice of $\hat{\mathbf{u}}(\hat{\mathbf{n}})$ and $\hat{\mathbf{v}}(\hat{\mathbf{n}})$. During tomographic reconstruction, whenever extracting a slice, we randomly choose $\hat{\mathbf{u}}(\hat{\mathbf{n}})$ in $\hat{\mathbf{n}}^\perp$ and $\hat{\mathbf{v}}(\hat{\mathbf{n}})$ in $\hat{\mathbf{u}}(\hat{\mathbf{n}})^\perp$. Theoretically, G (Equation (11)) is therefore a random variable. But we find that in practice, the impact of changing the seed is minimal, as shown in Figure 7. This could be attributed to the large number of angles L that are used for reconstruction and to g_θ probably being robust to rotations.



(a) Seed = 0 (b) Seed = 1 (c) Difference

Figure 7. Reconstructions for two different random seeds for a sample of the MosMed dataset.

A.2. Impact of the number of slices and angles

Figure 8 shows the impact of the number of angles L and the number of slices M on the reconstructions that are produced.

A.3. Dataset characteristics

Table 3 shows the number of samples used for training and validation in each dataset.

A.4. Dataset preprocessing details

Multiple Sclerosis In this longitudinal dataset, multiple studies (visits) are available for each patient. Each study includes T1, proton-density, T2, gadolinium-enhanced FLAIR and T1 weighted sequences, which were all rigidly registered to a common atlas and cropped. The resulting image shape is $54 \times 222 \times 179$. For a given study, we stack all 5 sequences along the channel axis to be input to the models, and apply channel-wise z-normalization. For each study, a ground truth segmentation mask was derived by a consensus of trained experts.

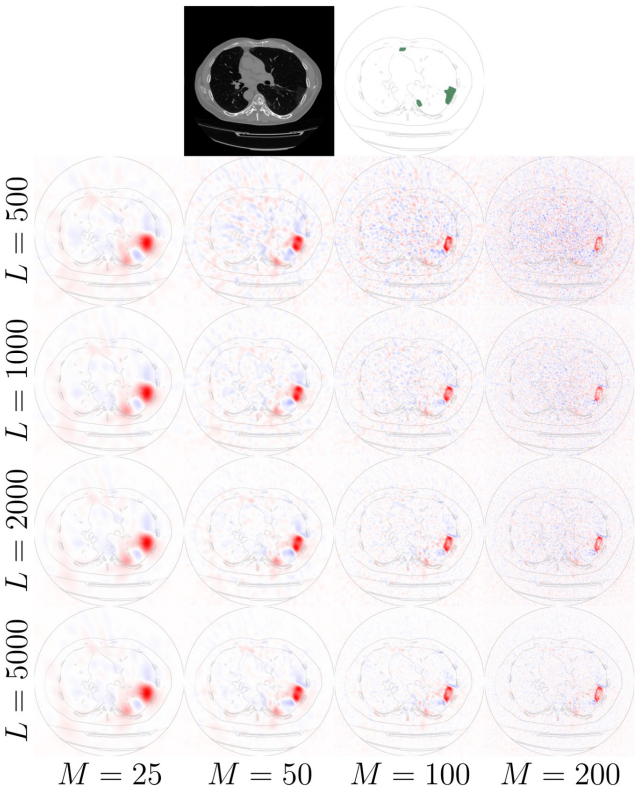


Figure 8. **ToNNO's output for different values of M and L** from Equations (8) and (9) for a sample of the MosMed dataset. The input image and ground truth segmentation are shown on top.

Dataset	Modality	Training		Validation	
		pos.	neg.	pos.	neg.
MS	MRI	1786	6363	181	783
AutoPET [15]	CT, PET	444	457	57	55
Duke [44]	MRI	975	699	92	66
MosMed [33]	CT	634	204	50	50

Table 3. Dataset characteristics and number of positive and negative samples used for training and validation.

AutoPET Each study consists of a CT image and an associated Standardized Uptake Value (SUV) image. In order to increase training speed, we cropped all images according to the following protocol: a binary mask was generated by thresholding the SUV image at 0.2. The minimum enclosing bounding box of all the positive voxels was then used to

crop the image. The CT scan was finally resampled onto this cropped image. We stack the CT and SUV images along the channel axis. The CT scan is divided by 1000 and the SUV by 10 at the input of the models. Ground truth manual segmentation masks are provided for each study.

MosMed There are 6 classes, named CT-0 to CT-6. CT-0 is the negative class, meaning that no signs of COVID-19 were identified in the scans, and CT-1 to 6 are the positive class and are sorted by increasing order of severity. We use only CT-0 and CT-1. The axial resolution of the CT scans is low, as only every tenth slice was kept in the public release of the dataset. All slices have shape 512×512. Ground truth semi-manual segmentation masks are available for 50 cases. They contain many tiny connected components, which is not suitable for our evaluation procedure. We therefore preprocessed these masks in order to reduce their number of connected components by applying a binary closing with a sphere of radius 20 followed by a binary opening with a sphere of radius 10. An example of this preprocessing is provided in Figure 9. We divide the volumes by 1000 at the input of the models.

(a) Input (b) Raw (c) Preprocessed

Figure 9. The masks of the MosMed datasets were preprocessed in order to remove tiny connected components.

Duke Each study comprises 6 different acquisitions: a T1-sequence, a fat-saturated pre-contrast T1 sequence, and 4 post-contrast T1 sequences. Trained experts provided bounding boxes delimiting the tumours. There is exactly one bounding-box per patient: even in the case of multiple tumours, only one was annotated. In order to obtain negative volumes, we separated the left and right breasts into two different images: the breast with bounding box was considered positive, the other breast was considered negative, unless the field *Contralateral Breast Involvement* was marked as positive in the clinical data, in which case the other breast was also considered positive. In this work, we only use the pre-contrast and first post-contrast sequences, as these were used by the human annotators to draw the bounding boxes. We stack them along the channel axis and apply channel-wise z-normalization. We manually verified that the coordinates of the bounding boxes aligned well with the tumours in all patients.

A.5. Detailed results for each model.

Table 4 shows results for each model, for each dataset.

Dataset	Model	F1-score	Dice/IoU	BA
Multiple Sclerosis	R-10-T, BN	0.75	0.41	0.93
	R-10-T, FBN	0.82	0.53	0.90
	R-10-T, GN	0.78	0.49	0.93
	R-10-T, GN, NP	0.21	0.20	0.52
	R-50, GN	**0.83**	**0.55**	**0.94**
AutoPET	R-10-T, BN	0.35	0.24	0.80
	R-10-T, FBN	0.39	**0.32**	0.74
	R-10-T, GN	0.36	0.27	0.77
	R-10-T, GN, NP	0.25	0.16	**0.81**
	R-50, GN	**0.42**	0.31	0.80
MosMed	R-10-T, BN	0.43	0.29	0.92
	R-10-T, FBN	0.50	0.35	**0.93**
	R-10-T, GN	0.50	0.35	0.92
	R-10-T, GN, NP	0.39	0.32	0.83
	R-50, GN	**0.54**	**0.39**	0.87
Duke	R-10-T, BN	0.38	0.30	0.71
	R-10-T, FBN	**0.51**	0.37	0.79
	R-10-T, GN	0.50	0.35	**0.80**
	R-10-T, GN, NP	0.27	**0.39**	0.54
	R-50, GN	0.45	0.36	0.77

Table 4. **Comparison of different model configurations**. BN: batch normalization, FBN: frozen batch normalization, GN: group normalization, NP: not pretrained.

A.6. Detailed training procedure

During a training step, we sample a batch of B samples $\{(V_{i_1}, y_{i_1}), ..., (V_{i_B}, y_{i_B})\}$ with replacement from our training dataset. For each sample, we sample a unit vector $\widehat{\mathbf{n}}$ from the uniform distribution over the unit sphere. We then extract M_{train} slices of shape 224×224 ($h_S = w_S = 224$) from that volume with normal vector $\widehat{\mathbf{n}}$, offsets s_m ranging from -1 to 1, and random vectors $\mathbf{u}_m, \mathbf{v}_m \in \widehat{\mathbf{n}}^{\perp}, m \in \{1, ..., M_{\text{train}}\}$. $\mathbf{u}_m$ and $\mathbf{v}_m$ are chosen perpendicular to each other. Their lengths are randomly and independently chosen in the range $[1 - 0.3, 1 + 0.3]$, which amounts to applying random anisotropic scale augmentations. Furthermore, we introduce slice-wise random translation in the range $(-0.3, 0.3)$ and affine intensity augmentations where the values of the pixels are shifted according to the function $y = ax + b$ with $a, b \sim \mathcal{U}(-0.3, 0.3)$. We concatenate the $B \times M_{\text{train}}$ slices in a batch and perform a training step by associating to each slice the label of the volume from which it came. In our experiments, we set $B = 2$ and $M_{\text{train}} = 120$. Furthermore, we always sample one positive and one negative volume per batch, as we observe that this improves learning. We use the

binary cross-entropy loss and optimize the neural network with Adam [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We use a cosine learning rate schedule [31] starting at $5 \cdot 10^{-5}$ and ending a 10^{-7}, which updates the learning rate at each training step. The number of training iterations for the Multiple Sclerosis, AutoPET, MosMed and Duke datasets are respectively set to 250,000, 187,500, 125,000 and 62,500. During training, we monitor the performance of the network on the validation set by obtaining volume-level predictions by max-pooling the slice level predictions. At the end of training, the validation balanced accuracies for the ResNet-10-T with frozen batch normalization where as follows: Multiple Sclerosis 0.93, AutoPET 0.88, MosMed 0.92, Duke 0.82.

A.7. Evaluation details

Heatmap binarization In order to compute segmentation metrics, we first need to binarize the heatmaps that we obtain with the different methods. We start by extracting the maximum value of each heatmap and find the dataset-dependent threshold τ that maximizes the weighted balanced accuracy:

$$wBA = \frac{5 \cdot \text{sensitivity} + \text{specificity}}{6}$$

where the predicted label for a sample is considered to be positive if the maximum value of the heatmap is above the threshold τ and negative otherwise. We give more weight to sensitivity because for medical applications, sensitivity is often more important than specificity: it may be more problematic for a tumour to go undetected than to detect something that is not a tumour. We tried to use this threshold τ to directly binarize the heatmaps, but this resulted in undersegmentation. We thus use a more advanced binarization procedure. First, we binarize the heatmaps with a threshold $\tau' < \tau$. Then, we extract each connected component, and reject the connected components whose maximum value in the heatmap is less than τ. This way, we obtain a better segmentation without changing the predicted binary label of a given sample. We grid-search the value of τ' that maximizes the dice score between the predicted and ground truth segmentations, using 5 ground truth segmentation masks per dataset. For the Duke dataset, which only has one bounding box annotation per positive sample, we pick the value of τ' such that the intersection over union (IoU) of the ground truth bounding box and the best matching predicted bounding box is maximized.

We chose not to optimize τ and τ' on the training set because the models risk being overfitted. We instead perform 10-fold Monte-Carlo cross-validation. For each fold, we optimize τ and τ' on the first half of the shuffled validation dataset (for optimizing τ' we randomly select 5 samples with ground truth mask/bounding box in this first half) and compute the metrics on the second half. The results that we report are averaged over the 10 folds.

Metrics For the Multiple Sclerosis, AutoPET and MosMed datasets, ground truth segmentations are available, which allows to compute the dice score with respect to the predicted segmentation. We then compute the connected components of the predicted and ground truth segmentations. We consider that a connected component in the prediction is a true positive if there is a connected component in the ground truth that has an IoU with it greater than $1/8$ (for reference, an IoU of $1/8$ corresponds to two overlapping spheres with respective radiuses r and $2r$). Otherwise, it is considered a false positive. Similarly, a connected component in the ground truth is considered a false negative if no connected component in the prediction has IoU with it greater than $1/8$. On each sample with non-empty ground truth segmentation, we then compute the precision, recall and F1-score. These three metrics, in addition to the dice score, are averaged over the different samples in the fold.

For the Duke breast cancer dataset, only one bounding box annotation per patient is available. We start by computing the connected components of the predicted segmentation. We then obtain the bounding boxes of the different connected components and compare them to the unique ground truth bounding box. We apply the same definitions of true positives, false positive and false negatives as above, using IoU between bounding boxes instead of IoU between segmentation masks. Instead of the dice score, we report the maximum IoU between the ground truth bounding box and any predicted bounding box. As only at most one bounding box is available for each patient, despite there sometimes being multiple tumours, many predicted connected components that are considered as false positives may actually be true positives. Thus, precision and F1-score are less relevant for this dataset.

For all datasets, we also report the global balanced accuracy, as defined in the previous paragraph, but with equal weights for sensitivity and specificity. We do not compute precision and F1-score on samples with empty ground truth segmentation. False positives in these samples thus have no effect on these metrics, but they are captured by the global balanced accuracy, which will be impacted if too many false positives are predicted for negative samples.

A.8. Results for additional baselines

In Table 5, we additionally provide results for ScoreCAM [55] and GradCAM++ [5]. We also provide results where the CAM methods are applied along all three spatial axes (i.e. axial, coronal and sagittal) and then averaged (denoted by a ★). As in the rest of the paper, all CAM methods were evaluated once for each layer (1 to 4) and the results presented in Table 5 are the best out of the four layers.

Dataset	Method	F1-score	Dice/Max IoU	Balanced accuracy
AutoPET	GradCAM	0.06	0.11	0.63
	GradCAM ★	0.05	0.15	0.65
	GradCAM++	0.22	0.21	0.61
	GradCAM++ ★	0.18	0.18	0.57
	LayerCAM	0.35	0.31	0.80
	LayerCAM ★	0.35	0.37	0.74
	ScoreCAM	0.23	0.23	0.63
	ScoreCAM ★	0.20	0.23	0.58
	ToNNO (ours)	0.39	0.32	0.74
	Averaged LayerCAM (ours)	0.40	**0.40**	**0.83**
	Tomographic LayerCAM (ours)	**0.49**	0.39	0.74
MosMed	GradCAM	0.23	0.24	0.78
	GradCAM ★	0.31	0.31	0.90
	GradCAM++	0.29	0.28	0.81
	GradCAM++ ★	0.26	0.38	0.89
	LayerCAM	0.39	0.35	0.90
	LayerCAM ★	0.48	0.44	0.84
	ScoreCAM	0.29	0.27	0.60
	ScoreCAM ★	0.38	0.38	0.85
	ToNNO (ours)	0.50	0.35	0.93
	Averaged LayerCAM (ours)	0.53	**0.48**	**0.95**
	Tomographic LayerCAM (ours)	**0.55**	0.41	0.89
Duke	GradCAM	0.07	0.15	0.51
	GradCAM ★	0.08	0.14	0.55
	GradCAM++	0.09	0.21	0.57
	GradCAM++ ★	0.34	0.30	**0.80**
	LayerCAM	0.24	0.29	0.58
	LayerCAM ★	0.44	0.38	0.74
	ScoreCAM	0.18	0.27	0.53
	ScoreCAM ★	0.23	0.20	0.60
	ToNNO (ours)	**0.51**	0.37	0.79
	Averaged LayerCAM (ours)	0.47	**0.42**	0.75
	Tomographic LayerCAM (ours)	**0.51**	**0.42**	0.79

Table 5. In this table, we provide results that were requested by the reviewers. Unfortunately, because of the premature termination of a contract with the company providing the private Multiple Sclerosis dataset, we had to delete all data and were unable to run new experiments on this dataset. ★ means averaging along the three directions (axial, coronal, sagittal).

A.9. Multiple Sclerosis dataset results

Method	Precision	Recall	F1-score	Dice	Balanced accuracy
GradCAM (layer 1)	0.04	0.27	0.06	0.04	0.57
GradCAM (layer 2)	0.15	0.38	0.18	0.12	0.64
GradCAM (layer 3)	0.03	0.02	0.02	0.06	0.86
GradCAM (layer 4)	0.00	0.00	0.00	0.02	0.91
LayerCAM (layer 1)	0.76	0.78	0.73	0.41	0.89
LayerCAM (layer 2)	0.51	0.50	0.48	0.24	0.91
LayerCAM (layer 3)	0.02	0.01	0.02	0.07	**0.94**
LayerCAM (layer 4)	0.00	0.00	0.00	0.02	0.91
ToNNO	0.82	0.88	0.82	0.53	0.90
Averaged LayerCAM (layer 1)	0.86	0.87	**0.84**	0.52	0.91
Averaged LayerCAM (layer 2)	0.66	0.70	0.65	0.30	0.90
Averaged LayerCAM (layer 3)	0.11	0.12	0.11	0.09	0.90
Averaged LayerCAM (layer 4)	0.15	0.14	0.14	0.07	0.91
Tomographic LayerCAM (layer 1)	0.77	0.86	0.78	0.55	0.88
Tomographic LayerCAM (layer 2)	**0.87**	0.87	**0.84**	**0.58**	**0.94**
Tomographic LayerCAM (layer 3)	0.83	**0.90**	**0.84**	0.57	0.93
Tomographic LayerCAM (layer 4)	0.81	0.87	0.81	0.53	0.91

Table 6. Results for the Multiple Sclerosis dataset.

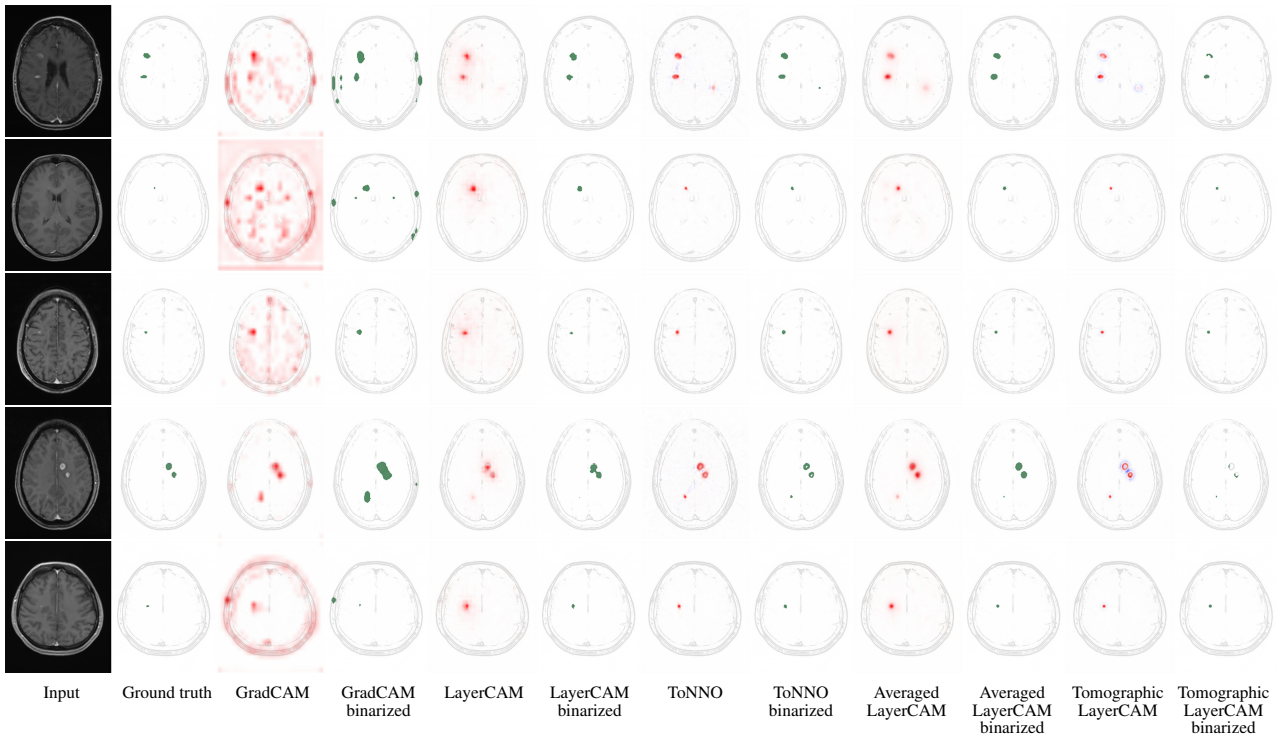


Figure 10. Examples for the Multiple Sclerosis dataset

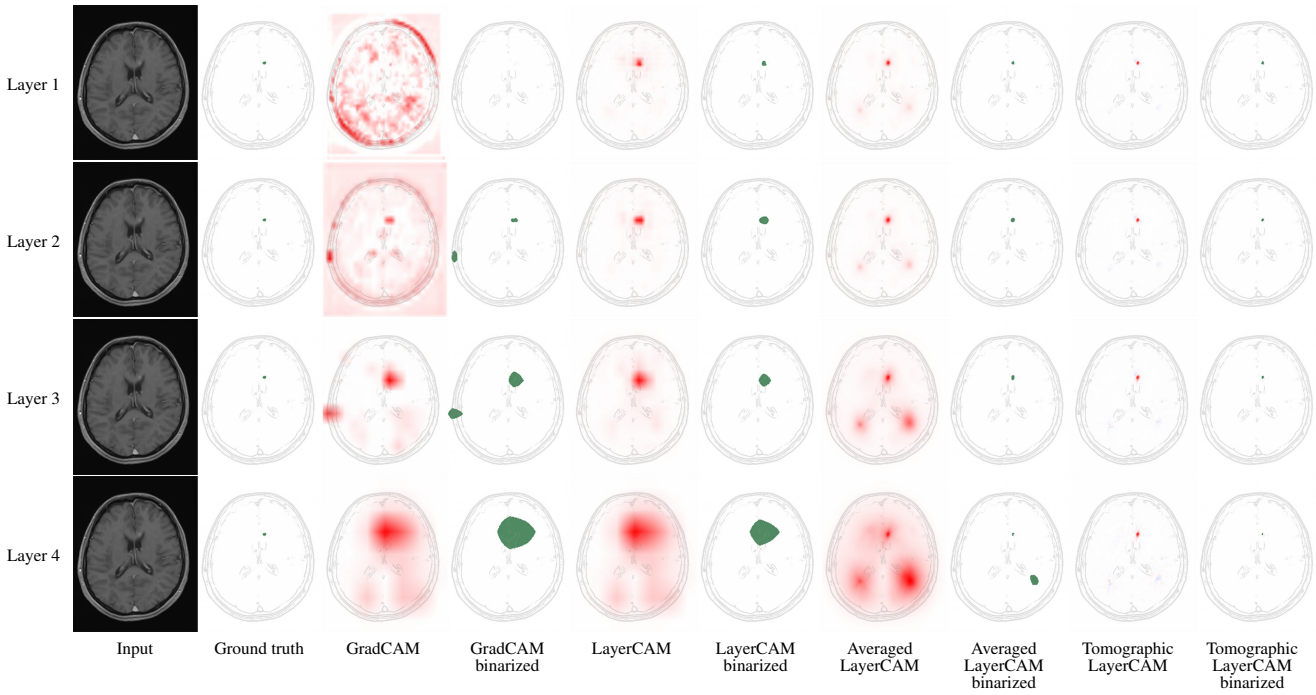


Figure 11. Output of GradCAM, LayerCAM, Averaged LayerCAM and Tomographic LayerCAM for layers 1 to 4 for one sample of the Multiple Sclerosis dataset.

A.10. AutoPET dataset results

Method	Precision	Recall	F1-score	Dice	Balanced accuracy
GradCAM (layer 1)	0.00	0.00	0.00	0.01	0.51
GradCAM (layer 2)	0.04	0.04	0.02	0.03	0.55
GradCAM (layer 3)	0.10	0.10	0.06	0.11	0.63
GradCAM (layer 4)	0.04	0.04	0.03	0.10	0.87
LayerCAM (layer 1)	0.43	0.30	0.29	0.28	0.69
LayerCAM (layer 2)	0.59	0.29	0.35	0.31	0.80
LayerCAM (layer 3)	0.26	0.14	0.16	0.20	0.84
LayerCAM (layer 4)	0.05	0.04	0.04	0.11	**0.89**
ToNNO	0.52	0.41	0.39	0.32	0.74
Averaged LayerCAM (layer 1)	0.68	0.34	0.38	0.37	0.82
Averaged LayerCAM (layer 2)	0.65	0.33	0.40	**0.40**	0.83
Averaged LayerCAM (layer 3)	0.44	0.21	0.25	0.24	0.81
Averaged LayerCAM (layer 4)	0.19	0.07	0.09	0.08	0.78
Tomographic LayerCAM (layer 1)	0.68	**0.47**	**0.49**	0.39	0.74
Tomographic LayerCAM (layer 2)	**0.71**	0.41	0.46	0.38	0.79
Tomographic LayerCAM (layer 3)	0.66	0.39	0.42	0.34	0.81
Tomographic LayerCAM (layer 4)	0.61	0.46	0.46	0.32	0.77

Table 7. Results for the AutoPET dataset.

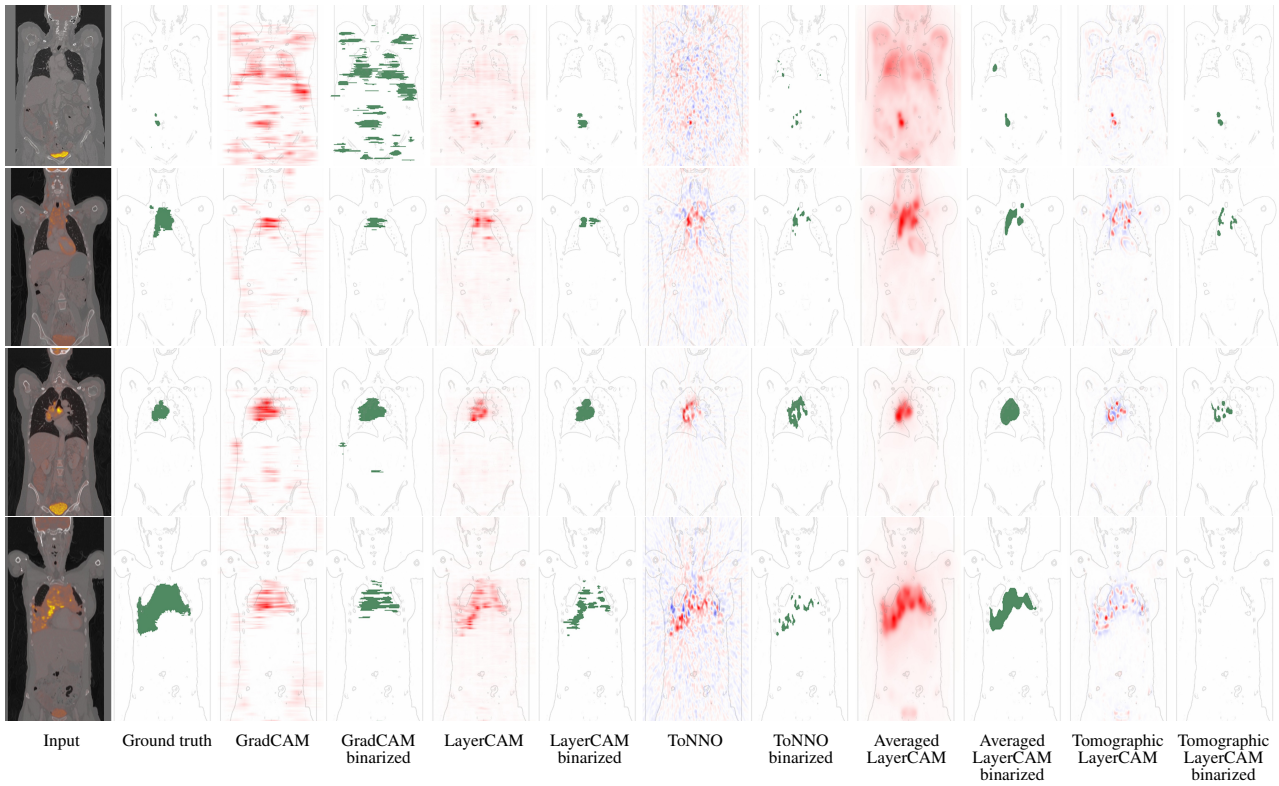


Figure 12. Examples for the AutoPET dataset

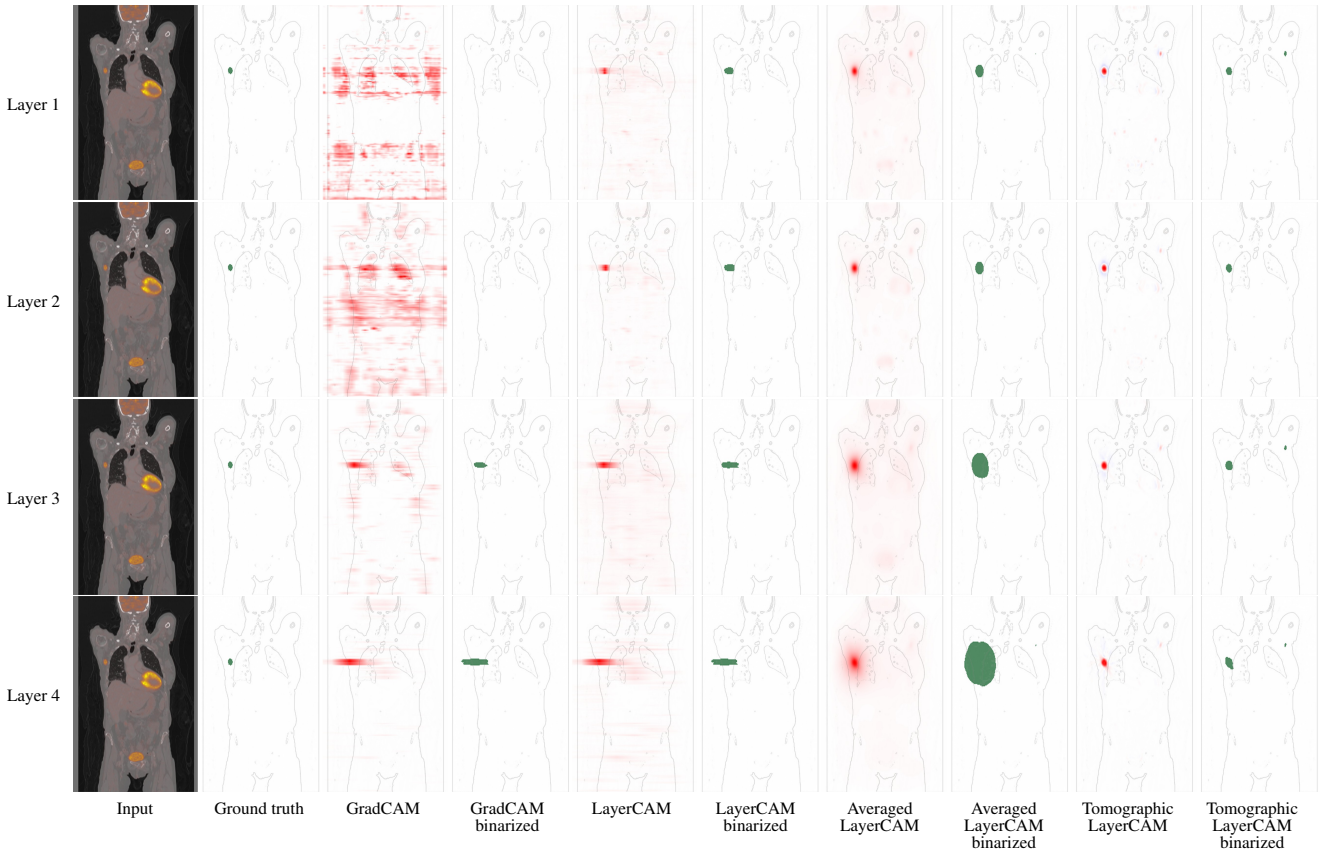


Figure 13. Output of GradCAM, LayerCAM, Averaged LayerCAM and Tomographic LayerCAM for layers 1 to 4 for one sample of the AutoPET dataset.

A.11. MosMed dataset results

Method	Precision	Recall	F1-score	Dice	Balanced accuracy
GradCAM (layer 1)	0.00	0.00	0.00	0.01	0.57
GradCAM (layer 2)	0.02	0.02	0.01	0.02	0.64
GradCAM (layer 3)	0.35	0.25	0.23	0.24	0.78
GradCAM (layer 4)	0.22	0.12	0.14	0.17	0.87
LayerCAM (layer 1)	0.52	0.25	0.29	0.28	0.85
LayerCAM (layer 2)	0.68	0.32	0.39	0.35	0.90
LayerCAM (layer 3)	0.45	0.28	0.30	0.28	0.78
LayerCAM (layer 4)	0.23	0.13	0.15	0.17	0.87
ToNNO	0.69	0.45	0.50	0.35	0.93
Averaged LayerCAM (layer 1)	0.70	0.44	0.49	0.41	0.90
Averaged LayerCAM (layer 2)	**0.72**	0.49	0.53	**0.48**	**0.95**
Averaged LayerCAM (layer 3)	0.68	0.37	0.43	0.37	**0.95**
Averaged LayerCAM (layer 4)	0.32	0.16	0.19	0.17	0.94
Tomographic LayerCAM (layer 1)	0.57	0.41	0.43	0.29	0.76
Tomographic LayerCAM (layer 2)	0.68	**0.52**	**0.55**	0.41	0.87
Tomographic LayerCAM (layer 3)	**0.72**	**0.52**	**0.55**	0.41	0.89
Tomographic LayerCAM (layer 4)	0.71	0.46	0.52	0.38	0.92

Table 8. Results for the MosMed dataset.

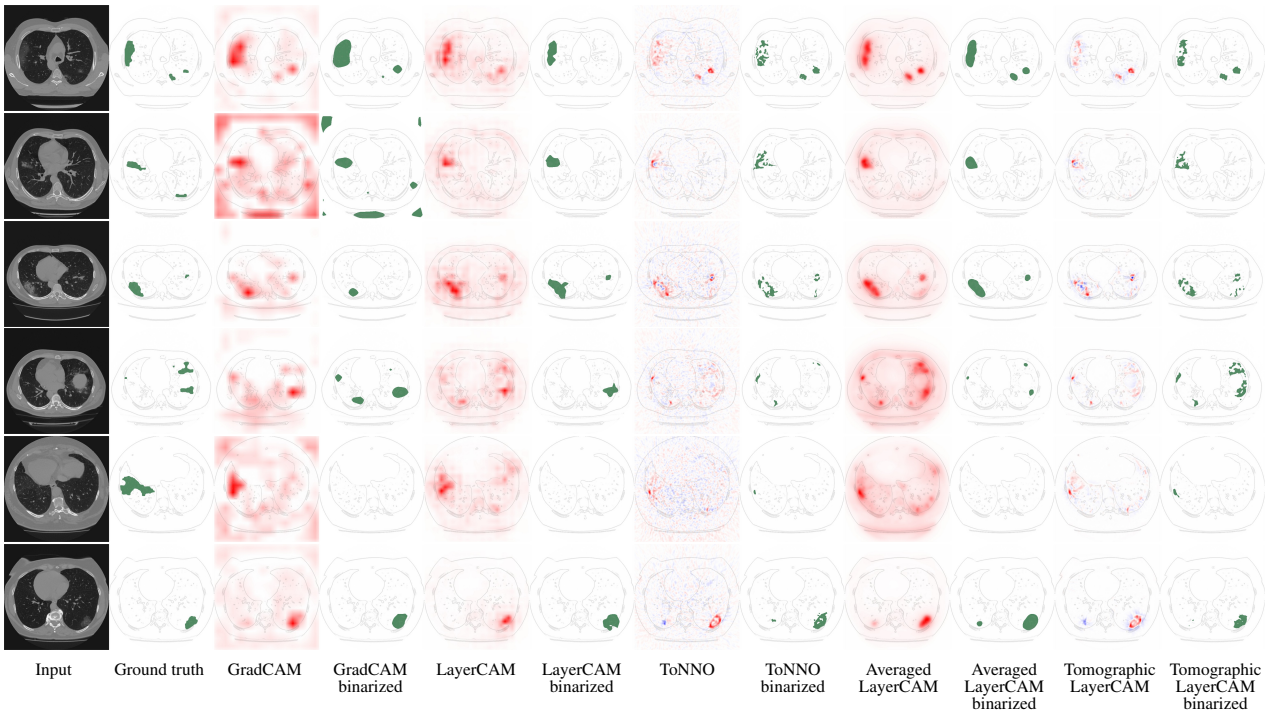


Figure 14. Examples for the MosMed COVID-19 thoracic CT dataset

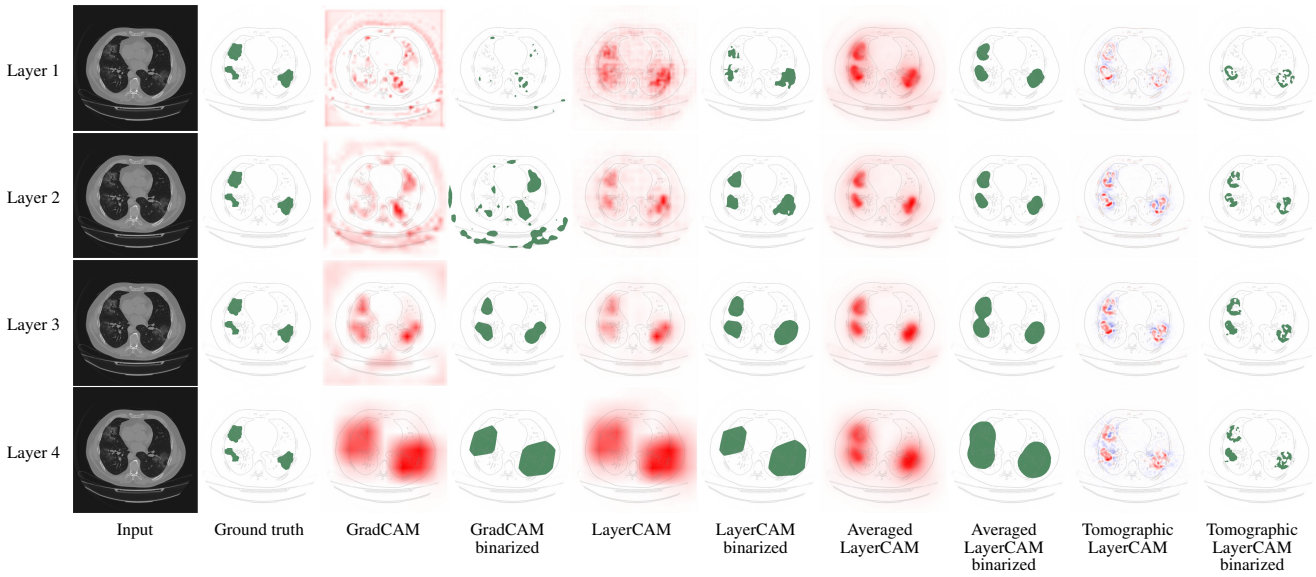


Figure 15. Output of GradCAM, LayerCAM, Averaged LayerCAM and Tomographic LayerCAM for layers 1 to 4 for one sample of the MosMed dataset.

A.12. Duke dataset results

Method	Precision	Recall	F1-score	MaxIoU	Balanced accuracy
GradCAM (layer 1)	0.00	0.07	0.01	0.03	0.54
GradCAM (layer 2)	0.01	0.11	0.01	0.04	0.52
GradCAM (layer 3)	0.04	0.53	0.07	0.15	0.51
GradCAM (layer 4)	0.09	0.22	0.11	0.08	0.65
LayerCAM (layer 1)	0.10	0.76	0.16	0.29	0.53
LayerCAM (layer 2)	0.17	0.81	0.24	0.29	0.58
LayerCAM (layer 3)	0.16	0.58	0.23	0.17	0.59
LayerCAM (layer 4)	0.11	0.25	0.13	0.08	0.65
ToNNO	**0.43**	0.77	**0.51**	0.37	0.79
Averaged LayerCAM (layer 1)	0.39	0.83	0.49	0.40	0.79
Averaged LayerCAM (layer 2)	0.37	**0.84**	0.47	**0.42**	0.75
Averaged LayerCAM (layer 3)	0.34	0.59	0.40	0.21	0.74
Averaged LayerCAM (layer 4)	0.16	0.22	0.18	0.09	0.71
Tomographic LayerCAM (layer 1)	0.15	0.82	0.23	0.39	0.55
Tomographic LayerCAM (layer 2)	0.30	0.83	0.40	0.40	0.68
Tomographic LayerCAM (layer 3)	0.41	0.81	**0.51**	**0.42**	0.79
Tomographic LayerCAM (layer 4)	0.42	0.80	**0.51**	0.39	**0.80**

Table 9. Results for the Duke dataset. Precision and F1-score are not as relevant as only at most one tumour per patient is reported in the ground truth data.

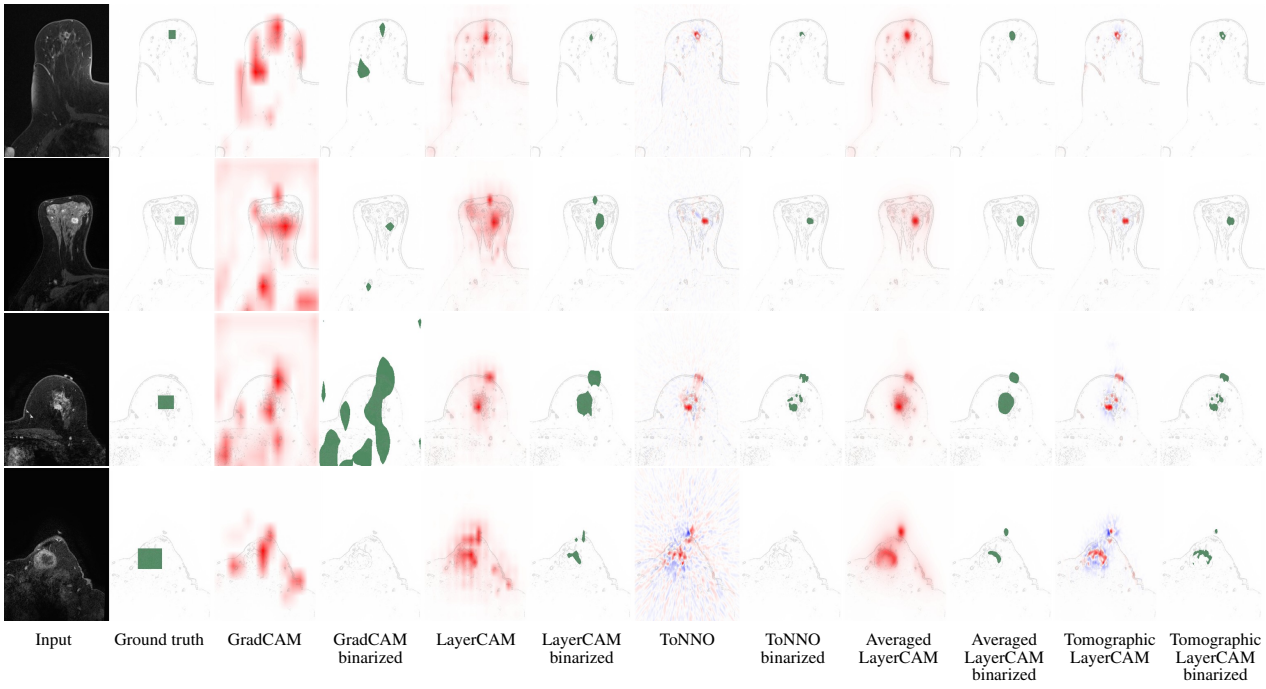


Figure 16. Examples for the Duke breast cancer MRI dataset

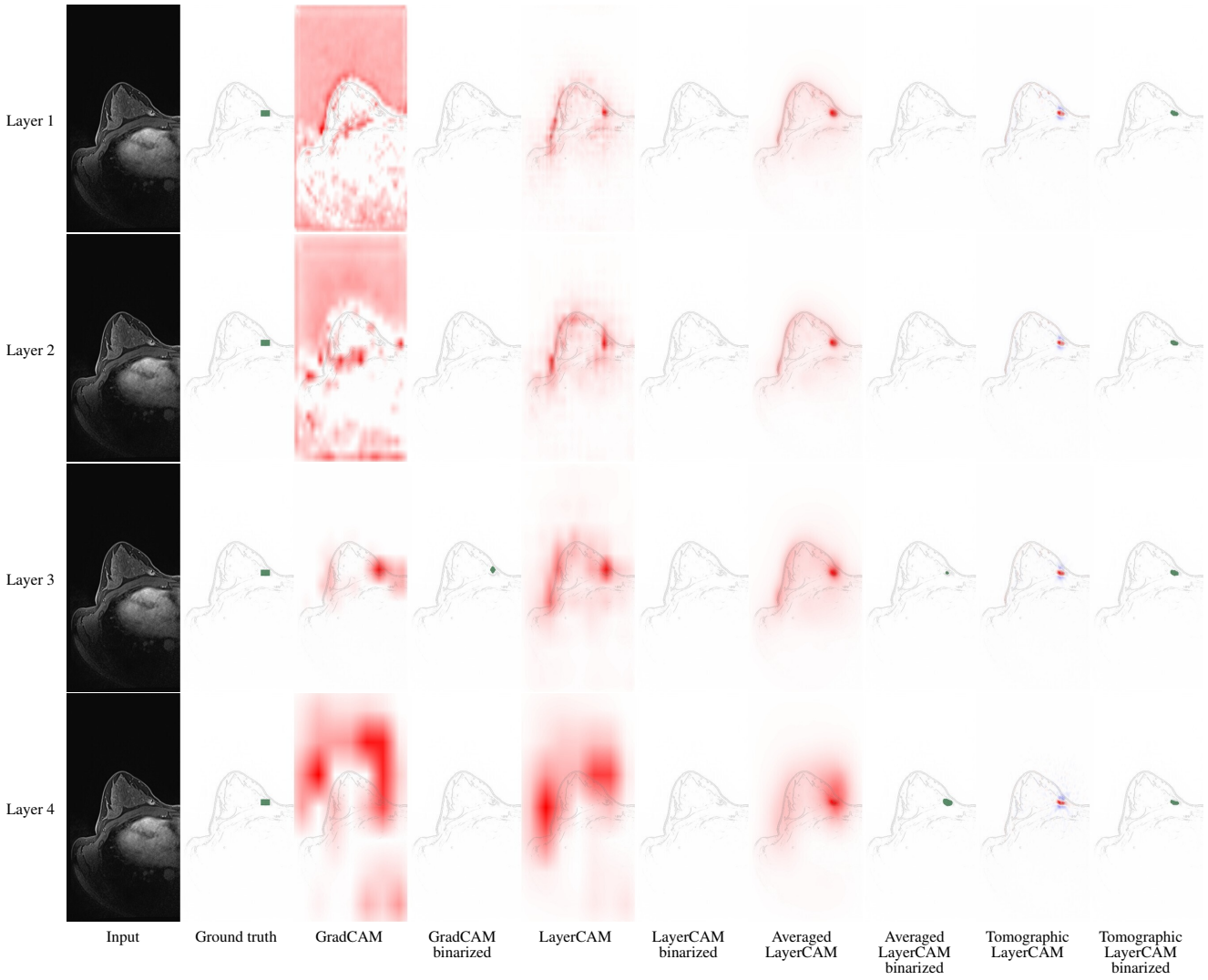


Figure 17. Output of GradCAM, LayerCAM, Averaged LayerCAM and Tomographic LayerCAM for layers 1 to 4 for one sample of the Duke dataset.