Aggregated Attributions for Explanatory Analysis of 3D Segmentation Models

Maciej Chrabaszcz[1,2] Hubert Baniecki[3] Piotr Komorowski[3]
Szymon Płotka[3,4] Przemyslaw Biecek[1,3]

[1]Warsaw University of Technology, Poland [2]NASK - National Research Institute, Poland
[3]University of Warsaw, Poland [4]University of Amsterdam, the Netherlands

A. Visualization of exemplary local explanations for different voxel attribution methods

See Figs. 1 to 5.

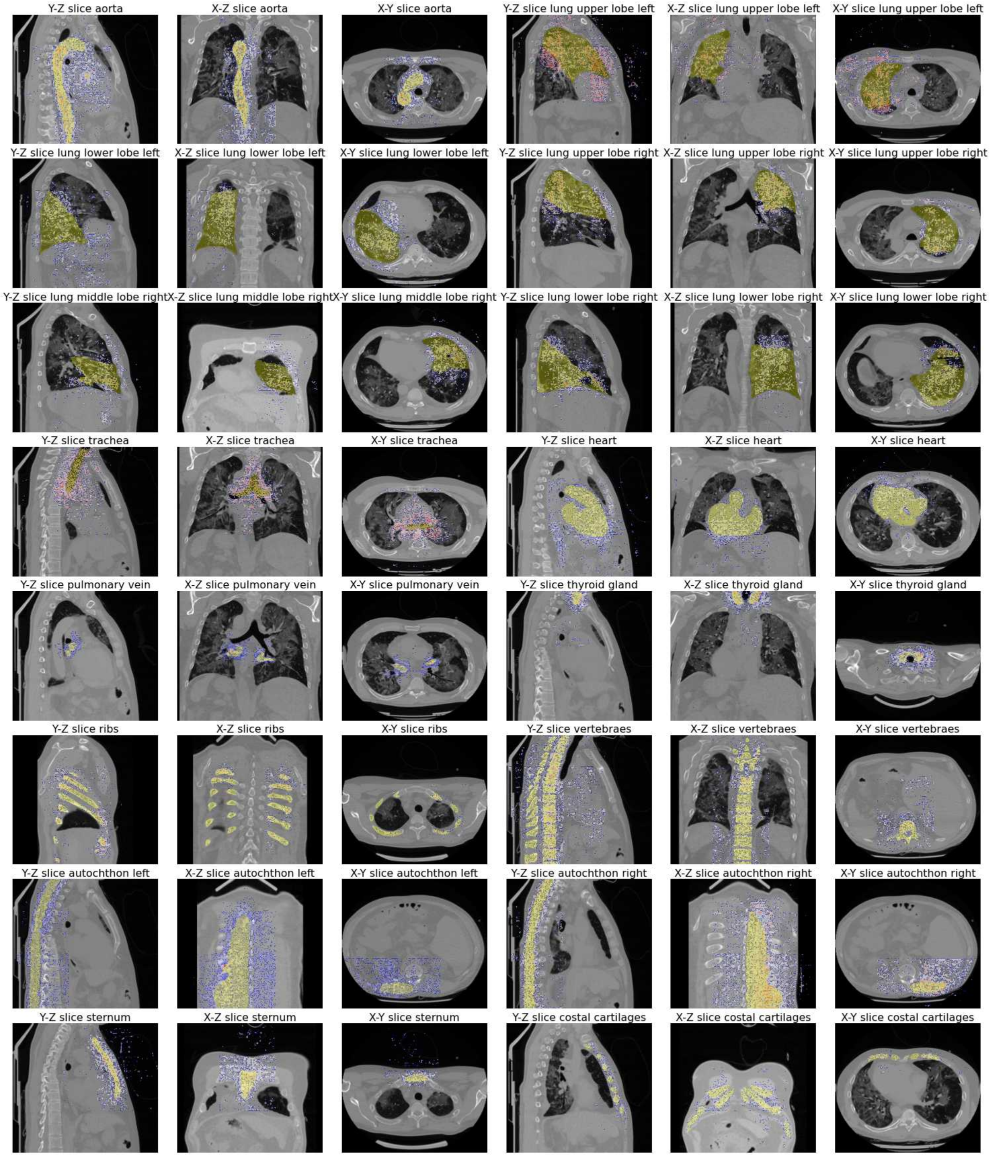
Y-Z slice aorta
X-Z slice aorta
X-Y slice aorta
Y-Z slice lung upper lobe left
X-Z slice lung upper lobe left
X-Y slice lung upper lobe left
Y-Z slice lung lower lobe left
X-Z slice lung lower lobe left
X-Y slice lung lower lobe left
Y-Z slice lung upper lobe right
X-Z slice lung upper lobe right
X-Y slice lung upper lobe right
Y-Z slice lung middle lobe right
X-Z slice lung middle lobe right
X-Y slice lung middle lobe right
Y-Z slice lung lower lobe right
X-Z slice lung lower lobe right
X-Y slice lung lower lobe right
Y-Z slice trachea
X-Z slice trachea
X-Y slice trachea
Y-Z slice heart
X-Z slice heart
X-Y slice heart
Y-Z slice pulmonary vein
X-Z slice pulmonary vein
X-Y slice pulmonary vein
Y-Z slice thyroid gland
X-Z slice thyroid gland
X-Y slice thyroid gland
Y-Z slice ribs
X-Z slice ribs
X-Y slice ribs
Y-Z slice vertebraes
X-Z slice vertebraes
X-Y slice vertebraes
Y-Z slice autochthon left
X-Z slice autochthon left
X-Y slice autochthon left
Y-Z slice autochthon right
X-Z slice autochthon right
X-Y slice autochthon right
Y-Z slice sternum
X-Z slice sternum
X-Y slice sternum
Y-Z slice costal cartilages
X-Z slice costal cartilages
X-Y slice costal cartilages

Figure 1. Visualization of VG attribution map across segmentation classes in three dimensions. The color mapping of attributions transitions from blue to red. Slices are chosen based on the highest area of segmented class within each dimension, with the top 95% values displayed for improved readability. Model predictions for individual organs are highlighted in yellow.

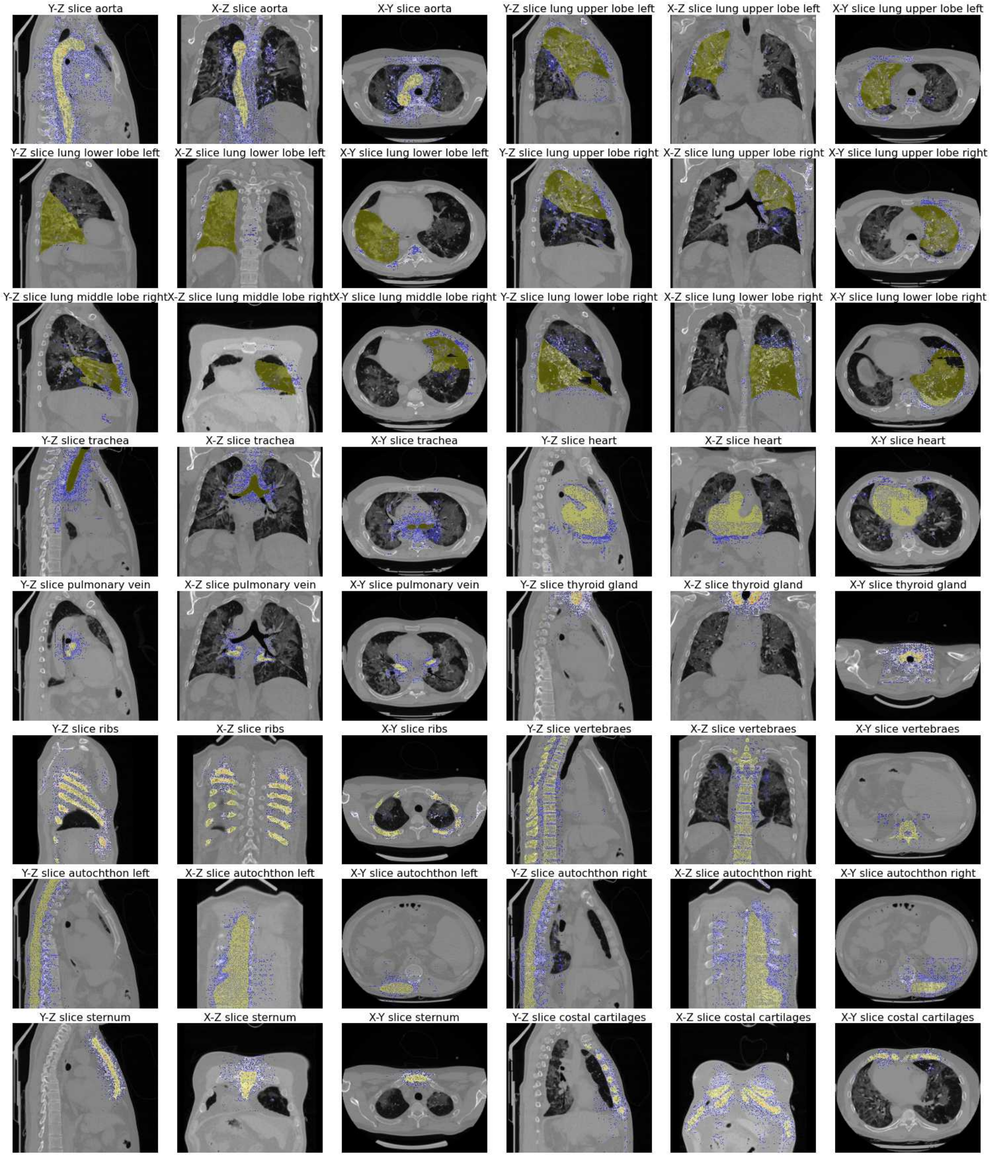
Y-Z slice aorta
X-Z slice aorta
X-Y slice aorta
Y-Z slice lung upper lobe left
X-Z slice lung upper lobe left
X-Y slice lung upper lobe left
Y-Z slice lung lower lobe left
X-Z slice lung lower lobe left
X-Y slice lung lower lobe left
Y-Z slice lung upper lobe right
X-Z slice lung upper lobe right
X-Y slice lung upper lobe right
Y-Z slice lung middle lobe right
X-Z slice lung middle lobe right
X-Y slice lung middle lobe right
Y-Z slice lung lower lobe right
X-Z slice lung lower lobe right
X-Y slice lung lower lobe right
Y-Z slice trachea
X-Z slice trachea
X-Y slice trachea
Y-Z slice heart
X-Z slice heart
X-Y slice heart
Y-Z slice pulmonary vein
X-Z slice pulmonary vein
X-Y slice pulmonary vein
Y-Z slice thyroid gland
X-Z slice thyroid gland
X-Y slice thyroid gland
Y-Z slice ribs
X-Z slice ribs
X-Y slice ribs
Y-Z slice vertebraes
X-Z slice vertebraes
X-Y slice vertebraes
Y-Z slice autochthon left
X-Z slice autochthon left
X-Y slice autochthon left
Y-Z slice autochthon right
X-Z slice autochthon right
X-Y slice autochthon right
Y-Z slice sternum
X-Z slice sternum
X-Y slice sternum
Y-Z slice costal cartilages
X-Z slice costal cartilages
X-Y slice costal cartilages

Figure 2. Visualization of IG attribution map across segmentation classes in three dimensions. The color mapping of attributions transitions from blue to red. Slices are chosen based on the highest area of segmented class within each dimension, with the top 95% values displayed for improved readability. Model predictions for individual organs are highlighted in yellow.

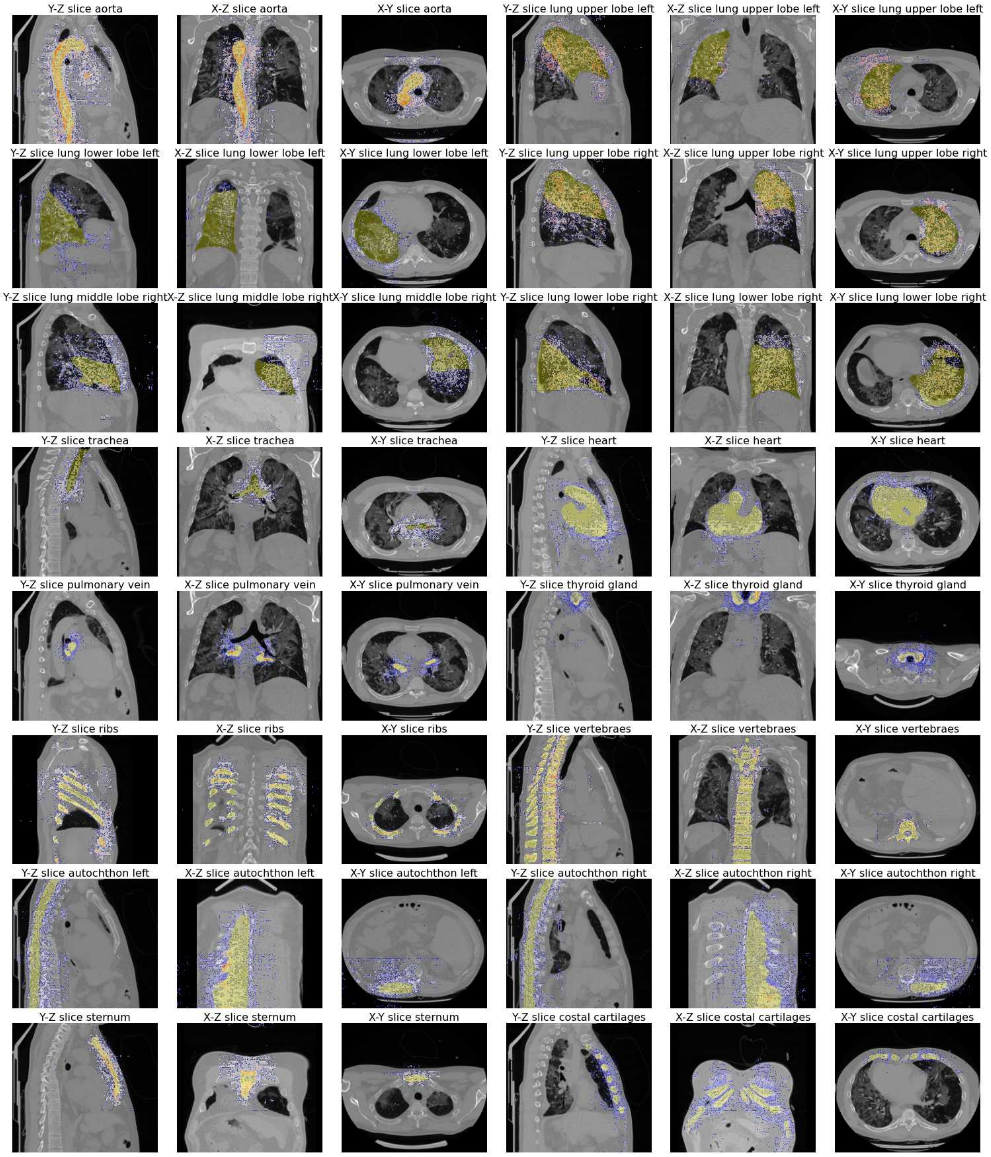
Y-Z slice aorta
X-Z slice aorta
X-Y slice aorta
Y-Z slice lung upper lobe left
X-Z slice lung upper lobe left
X-Y slice lung upper lobe left
Y-Z slice lung lower lobe left
X-Z slice lung lower lobe left
X-Y slice lung lower lobe left
Y-Z slice lung upper lobe right
X-Z slice lung upper lobe right
X-Y slice lung upper lobe right
Y-Z slice lung middle lobe right
X-Z slice lung middle lobe right
X-Y slice lung middle lobe right
Y-Z slice lung lower lobe right
X-Z slice lung lower lobe right
X-Y slice lung lower lobe right
Y-Z slice trachea
X-Z slice trachea
X-Y slice trachea
Y-Z slice heart
X-Z slice heart
X-Y slice heart
Y-Z slice pulmonary vein
X-Z slice pulmonary vein
X-Y slice pulmonary vein
Y-Z slice thyroid gland
X-Z slice thyroid gland
X-Y slice thyroid gland
Y-Z slice ribs
X-Z slice ribs
X-Y slice ribs
Y-Z slice vertebraes
X-Z slice vertebraes
X-Y slice vertebraes
Y-Z slice autochthon left
X-Z slice autochthon left
X-Y slice autochthon left
Y-Z slice autochthon right
X-Z slice autochthon right
X-Y slice autochthon right
Y-Z slice sternum
X-Z slice sternum
X-Y slice sternum
Y-Z slice costal cartilages
X-Z slice costal cartilages
X-Y slice costal cartilages

Figure 3. Visualization of SG attribution map across segmentation classes in three dimensions. The color mapping of attributions transitions from blue to red. Slices are chosen based on the highest area of segmented class within each dimension, with the top 95% values displayed for improved readability. Model predictions for individual organs are highlighted in yellow.

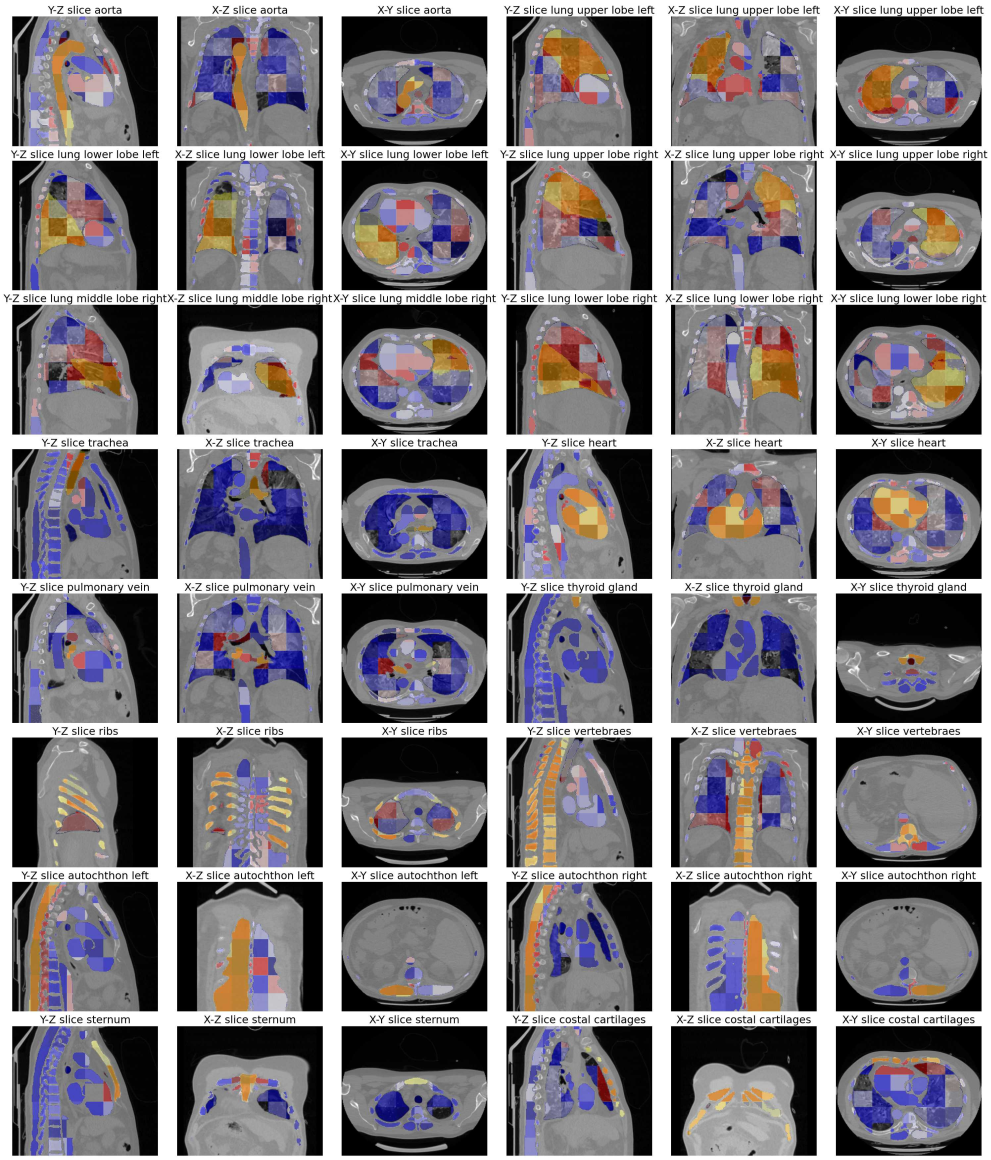
Y-Z slice aorta
X-Z slice aorta
X-Y slice aorta
Y-Z slice lung upper lobe left
X-Z slice lung upper lobe left
X-Y slice lung upper lobe left
Y-Z slice lung lower lobe left
X-Z slice lung lower lobe left
X-Y slice lung lower lobe left
Y-Z slice lung upper lobe right
X-Z slice lung upper lobe right
X-Y slice lung upper lobe right
Y-Z slice lung middle lobe right
X-Z slice lung middle lobe right
X-Y slice lung middle lobe right
Y-Z slice lung lower lobe right
X-Z slice lung lower lobe right
X-Y slice lung lower lobe right
Y-Z slice trachea
X-Z slice trachea
X-Y slice trachea
Y-Z slice heart
X-Z slice heart
X-Y slice heart
Y-Z slice pulmonary vein
X-Z slice pulmonary vein
X-Y slice pulmonary vein
Y-Z slice thyroid gland
X-Z slice thyroid gland
X-Y slice thyroid gland
Y-Z slice ribs
X-Z slice ribs
X-Y slice ribs
Y-Z slice vertebraes
X-Z slice vertebraes
X-Y slice vertebraes
Y-Z slice autochthon left
X-Z slice autochthon left
X-Y slice autochthon left
Y-Z slice autochthon right
X-Z slice autochthon right
X-Y slice autochthon right
Y-Z slice sternum
X-Z slice sternum
X-Y slice sternum
Y-Z slice costal cartilages
X-Z slice costal cartilages
X-Y slice costal cartilages

Figure 4. Visualization of KernelSHAP (cubes) attribution map across segmentation classes in three dimensions. The color mapping of attributions transitions from blue to red, attributions are zeroed when they overlap with the background class for better readability. Slices are chosen based on the highest area of segmented class within each dimension. Model predictions for individual organs are highlighted in yellow.

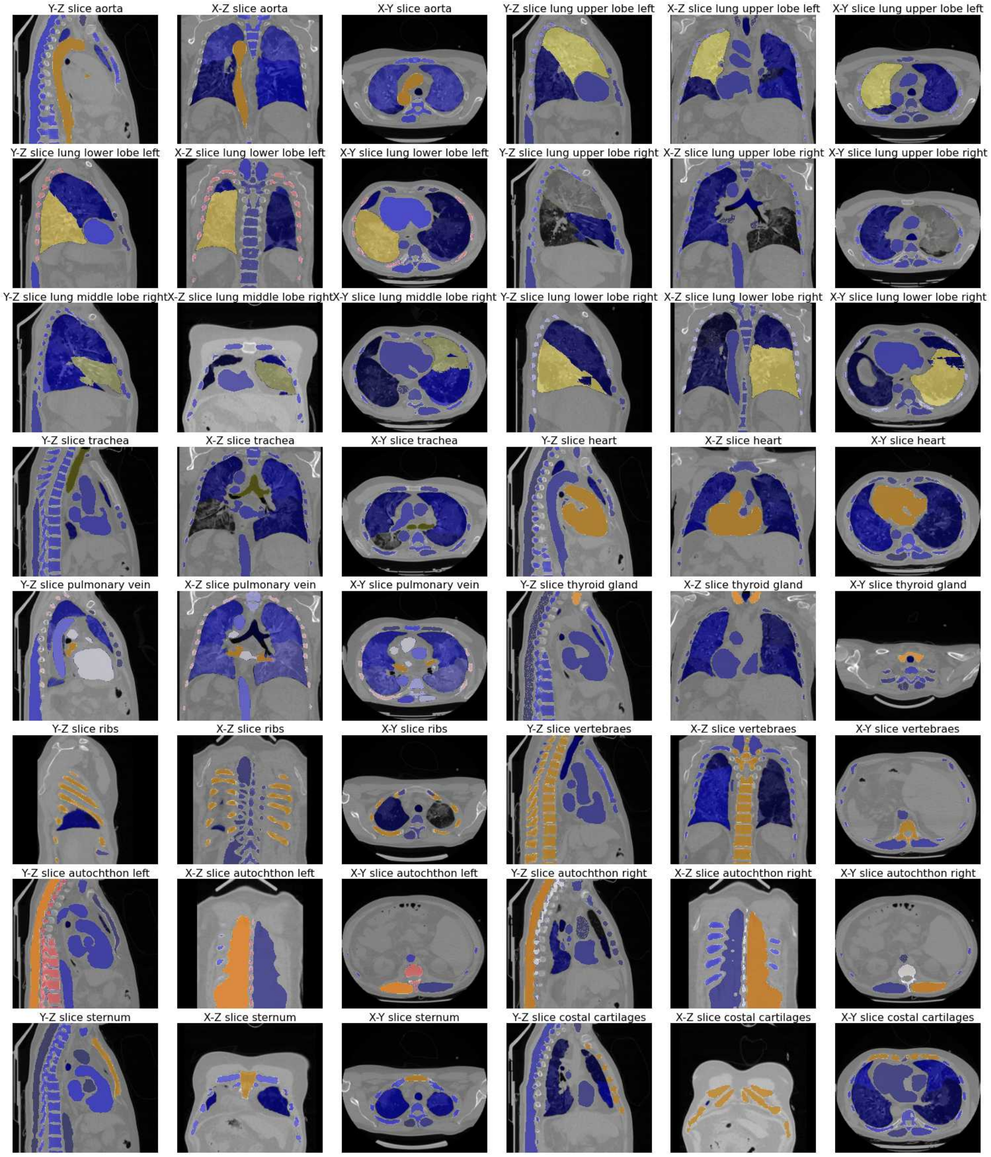
Y-Z slice aorta
X-Z slice aorta
X-Y slice aorta
Y-Z slice lung upper lobe left
X-Z slice lung upper lobe left
X-Y slice lung upper lobe left
Y-Z slice lung lower lobe left
X-Z slice lung lower lobe left
X-Y slice lung lower lobe left
Y-Z slice lung upper lobe right
X-Z slice lung upper lobe right
X-Y slice lung upper lobe right
Y-Z slice lung middle lobe right
X-Z slice lung middle lobe right
X-Y slice lung middle lobe right
Y-Z slice lung lower lobe right
X-Z slice lung lower lobe right
X-Y slice lung lower lobe right
Y-Z slice trachea
X-Z slice trachea
X-Y slice trachea
Y-Z slice heart
X-Z slice heart
X-Y slice heart
Y-Z slice pulmonary vein
X-Z slice pulmonary vein
X-Y slice pulmonary vein
Y-Z slice thyroid gland
X-Z slice thyroid gland
X-Y slice thyroid gland
Y-Z slice ribs
X-Z slice ribs
X-Y slice ribs
Y-Z slice vertebraes
X-Z slice vertebraes
X-Y slice vertebraes
Y-Z slice autochthon left
X-Z slice autochthon left
X-Y slice autochthon left
Y-Z slice autochthon right
X-Z slice autochthon right
X-Y slice autochthon right
Y-Z slice sternum
X-Z slice sternum
X-Y slice sternum
Y-Z slice costal cartilages
X-Z slice costal cartilages
X-Y slice costal cartilages

Figure 5. Visualization of KernelSHAP (semantic) attribution map across segmentation classes in three dimensions. The color mapping of attributions transitions from blue to red, attributions are zeroed when they overlap with background class for better readability. Slices are chosen based on the highest area of segmented class within each dimension. Model predictions for individual organs are highlighted in yellow.

B. Experimental setup: hyperparameters

B.1. Attribution methods parameters

- **Integrated Gradients** We used $x' = 0$ as baseline and $n = 20$ while calculating IG attribution.
- **SmoothGrad** We used $n = 20$ and gaussian noise with $\sigma = 0.1 \cdot (\max(x) - \min(x))$.
- **Kernel SHAP (cubes)** We used 512 components for cubes and 1000 samples for attribution calculation. We replace selected features (cubes) with zeros.
- **Kernel SHAP (semantic)** We used 200 samples and replaced features (segmentation regions) with zeros.

B.2. Attributions Evaluations metrics parameters

When calculating each metric we normalized attributions into $[0, 1]$.

- **Faithfulness** We used subset size $S = 224^2$. We used Pearson's Correlation as a similarity function. We replaced selected regions with zeros. We used $n = 100$.
- **Sensitivity** For sensitivity, we used $n = 3$ due to the computational complexity of Kernel SHAP attributions and a lower bound of 0.1.
- **Complexity** When calculating Complexity we used $\theta = 0.1$.

B.3. Implementation details of baseline model for 3D Image Segmentation

The model is trained using $2\times$ NVIDIA A100 40GB GPUs. As input, we use a patch of $96 \times 96 \times 96$ with a batch size of 4. We employ an AdamW optimizer with an initial learning rate of $1\mathrm{e}{-4}$ and weight decay of $1\mathrm{e}{-5}$ to minimize the loss function $\mathcal{L}$, which is defined as:

$$\mathcal{L} = \mathcal{L}_D + \Lambda \cdot \mathcal{L}_{CE}, \tag{1}$$

where $\mathcal{L}_D$, $\mathcal{L}_{CE}$ are Dice and Cross-Entropy loss, respectively. A grid search optimization ∈ [0.5, 1.0] was performed which estimated an optimal value $\Lambda = 1$.

We use a cosine annealing learning rate scheduler [5]. We implemented our network in Python v3.9 using PyTorch v2.1 and the MONAI library [1]. As an evaluation metric, we use the Dice Similarity Coefficient. We employ a one-way analysis of variance to evaluate the significance of variations among the segmentation performance metrics. We use $p < 0.05$ to distinguish a statistically significant difference.

C. Use-case: explanatory analysis of a model for segmenting anatomical structures in CT scans

Here, we provide supplementary visualizations of explanations computed for both datasets. In Fig. 6, we summarize the AGG^{2}EXP methodology with four explanation visualizations. Fig. 7 shows the distribution of local RoI importances for TSV2; analogous to **??** for B50. Fig. 8 shows the global RoI importance on a graph for B50 analogous to **??** for TSV2. Note that there are no annotations of lung pathologies available in the TSV2 dataset. We acknowledge that explanations computed on the external B50 dataset have more outliers, which seems reasonable as the model was trained on TSV2. Moreover, especially comparing Fig. 8 to Fig. 6 (**bottom-right**), we observe differences in how the model utilizes the contextual physical and semantical information between the two validation datasets.

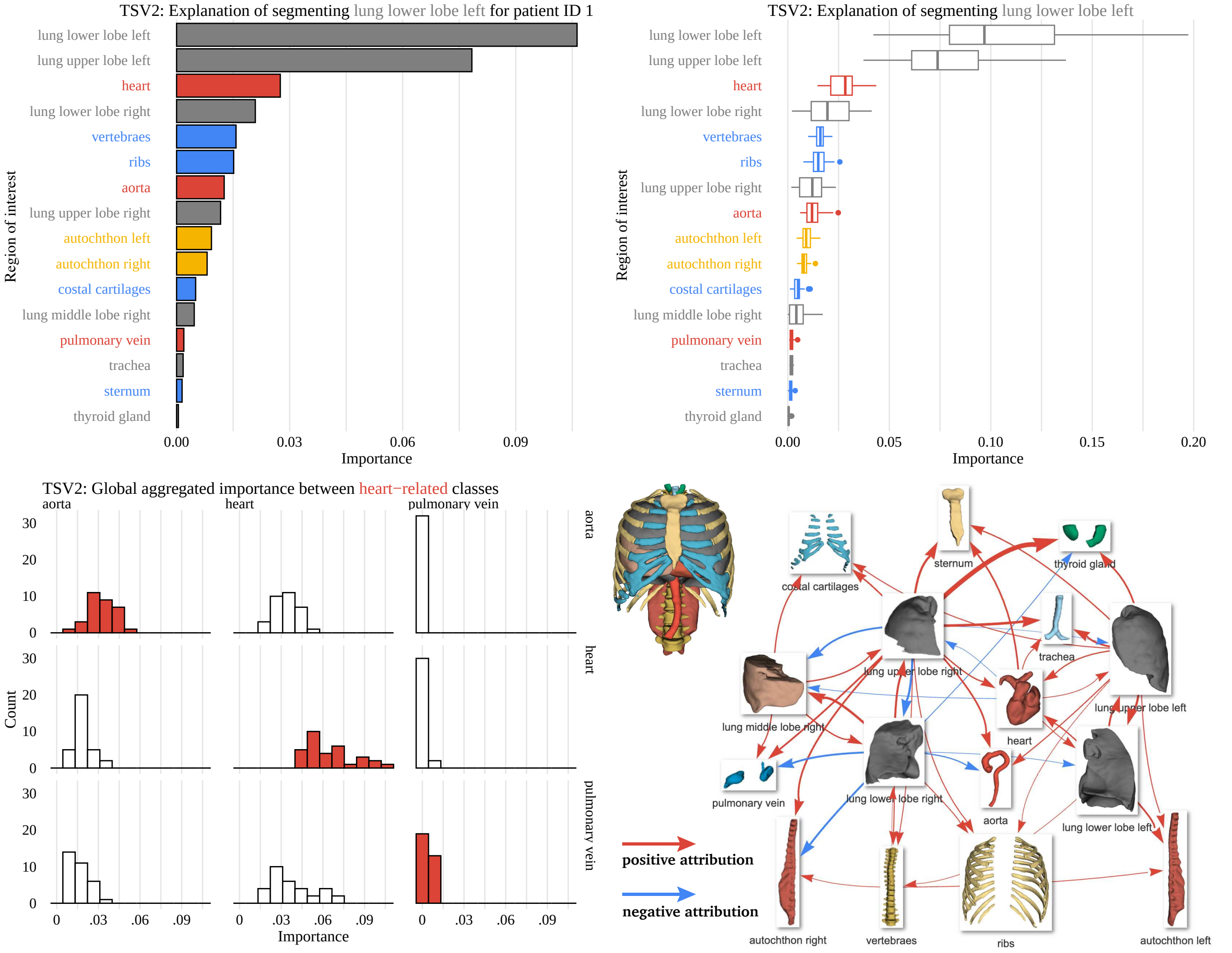

Figure 6. AGG^{2}EXP aggregates attributions into a global RoI importances denoting what a 3D segmentation model has learned. **Top-left:** We aggregate local voxel attributions into local RoI importance for a single input (patient) and a single output class (*lung lower lobe left*). **Top-right:** We aggregate local RoI importances for a subset of inputs (patients) into global RoI importance. **Bottom-left:** Global analysis of RoI importances for a subset of inputs between pairs of heart-related class labels. **Bottom-right:** AGG^{2}EXP allows to discover a higher-level representation acquired by the complex segmentation model. Explanations are aggregated twice to obtain global importance between output class labels and other potential RoIs, e.g. lung pathologies.

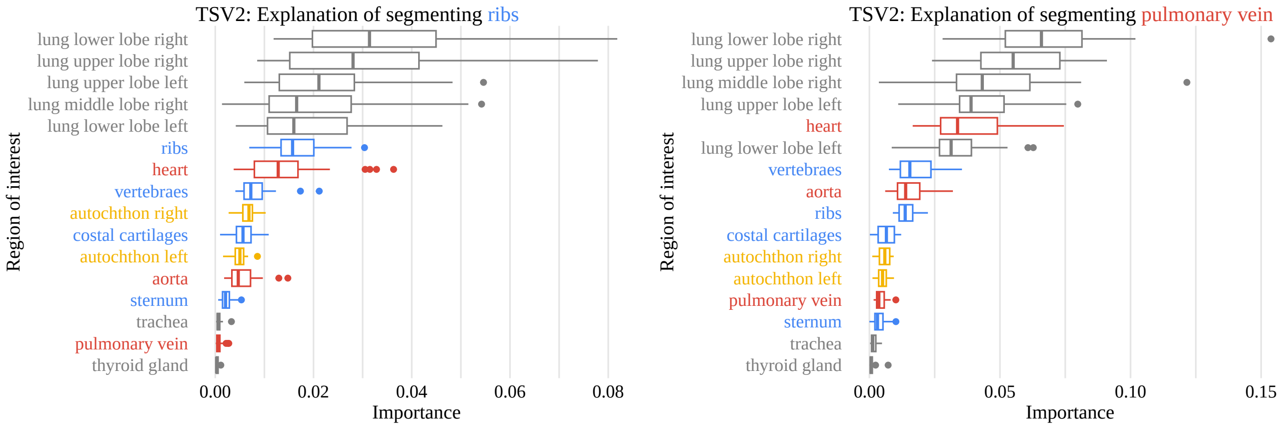

Figure 7. Distribution of local RoI importances for two class labels: *ribs* and *pulmonary vein*. We color objects by their semantic meaning, i.e., cardiovascular system in **red**, muscles in **yellow**, bones in **blue**, other organs in **grey**.

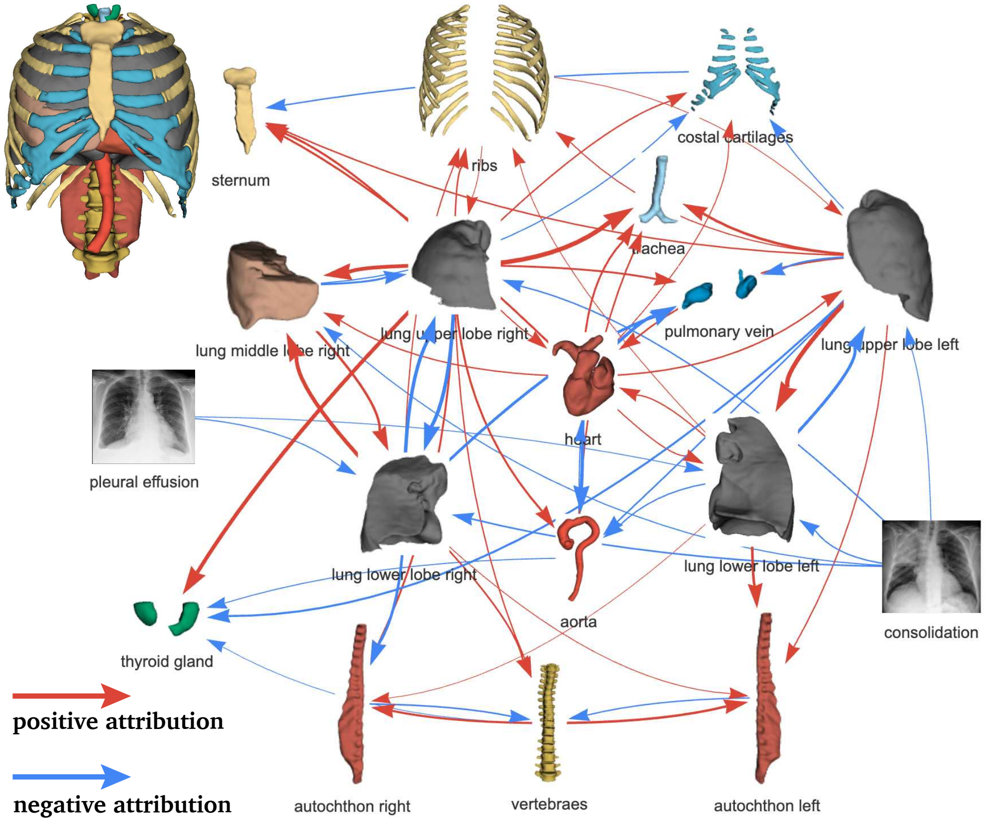

Figure 8. A global RoI importance explanation of the model's 3D segmentation. Aggregated attributions provide a measure of semantic importance between the segmented objects in B50. We visualize about one-third of the most important edges for clarity. The X-rays portraying the regions of lung pathologies are for illustration purposes only.

C.1. Analysis of outliers with AGG^{2}EXP

Tests' p-values and correlation coefficients are available in Tab. 1. When training Isolation Forest on all labels, we used 100 estimators. Each IF model was trained on the TSV2 train set that includes 490 CT scans.

Table 1. P-values and correlation coefficients from Spearman correlation test between Dice and anomaly scores of the IF models.

Label	**p-value**	**Spearman Correlation**
lung lower lobe right	0.000194	0.613
lung upper lobe right	0.000354	0.592
background	0.000432	0.585
costal cartilages	0.00431	0.491
lung middle lobe right	0.0083	0.459
heart	0.0181	0.415
sternum	0.0324	0.379
autochthon left	0.0399	0.365
ribs	0.129	0.274
lung upper lobe left	0.13	0.273
autochthon right	0.139	0.267
trachea	0.145	0.264
lung lower lobe left	0.306	0.187
aorta	0.446	-0.14
thyroid gland	0.463	-0.135
pulmonary vein	0.55	0.11
vertebraes	0.842	0.0367

D. Comparison of the segmentation performance between the baseline models

Table 2. TotalSegmentator test set performance comparison of mean Dice Score Coefficient (mDSC) for the following classes. For better readability, we present the left and right lungs as means of their subparts and autochthon as one class: Aorta (A), Left Lung (LL), Right Lung (RL), Trachea (T), Heart (H), Pulmonary Vein (PV), Thyroid Gland (TG), Ribs (R), Vertebrae (V), Autochthon (AU), Sternum (S), Costal Cartilages (CC). *indicates statistical significance between Swin UNETRv2 and other state-of-the-art methods mDSC ($p < 0.05$). **indicates training with self-supervised pre-trained weights.

Method	**A**	**LL**	**RL**	**T**	**H**	**PV**	**TG**	**R**	**V**	**AU**	**S**	**CC**	**mDSC ↑**
Swin UNETRv1** [6]	78.02	85.50	83.93	85.36	87.93	69.26	65.23	82.07	84.68	83.20	63.57	75.42	82.16 (*)
3D U-Net [2]	86.35	90.80	88.26	85.15	90.57	71.91	71.81	89.19	86.56	88.44	81.73	78.65	87.11 (*)
UNETR [3]	87.23	90.27	87.94	87.20	90.01	74.65	72.85	91.90	89.65	85.45	81.70	81.48	87.41 (*)
Swin UNETRv1 [6]	89.13	91.75	88.62	89.27	90.08	78.51	78.88	92.03	92.25	90.32	80.74	85.13	89.38 (*)
Swin UNETRv2 [4]	89.15	92.80	90.34	89.17	91.18	77.81	80.41	92.28	92.65	91.35	83.16	85.80	**90.24**

References

[1] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 7

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 10

[3] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 10

[4] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 416–426. Springer, 2023. 10

[5] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 7

[6] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 10