Surface-Conditioned Implicit Reconstruction of Internal Anatomy for Automated Patient Positioning

Denis Krnjaca[0009−0009−5163−3282] and Mattias P. Heinrich[0000−0002−7489−1972]

Institute of Medical Informatics, University of Lübeck, Lübeck, Germany
denis.krnjaca@uni-luebeck.de

Abstract. Automated pre-scan positioning can improve radiology workflow efficiency by reducing manual interaction during patient setup and scan planning. While external body surfaces can be acquired non-invasively using RGB-D sensors, organ-specific scan planning requires information about internal anatomy that is not directly observable. We formulate internal anatomy estimation from body-surface point clouds as a surface-conditioned implicit reconstruction problem. Our model combines a hierarchical Point Transformer encoder with a modulated SIREN decoder. The encoder extracts a patient-specific latent representation from the body surface, while the decoder predicts anatomical class logits at arbitrary 3D query coordinates, enabling dense volumetric reconstruction without a fixed output grid. We train on 4,780 subjects and evaluate on an independent test set of 1,454 subjects from the German National Cohort (NAKO), reconstructing 20 internal anatomical structures from MRI-derived body-surface point clouds simulating depth observations. Compared with the point-set-based DeformingPointTransformer (DPT) baseline, our method achieves a mean ASSD of 2.95 mm versus the reported DPT mean CD of 4.78 mm. For HD95, reported for both methods, the aggregate error decreases from 12.06 mm to 9.26 mm, with lower values for 17 of 20 structures. A qualitative zero-shot feasibility experiment on ten unlabeled real depth observations further shows that the trained model can process sensor-derived body surfaces without fine-tuning. These results support surface-conditioned implicit representations for dense internal anatomy estimation from external body observations. Code: https://github.com/multimodallearning/implicit-anatomy

Keywords: Implicit neural representations · Multi-organ segmentation · Automated patient positioning.

1 Introduction

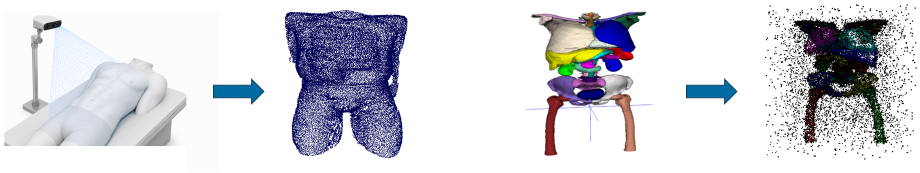
Diagnostic imaging quality and efficiency depend strongly on preparatory steps before acquisition, including patient positioning, table centering, scan range selection, and scan-specific parameter planning [11]. Current CT and MRI workflows support these steps with scout or localizer scans; however, manual interaction and operator experience remain crucial [17]. Inaccurate positioning or

suboptimal planning can lead to repeated scans, longer preparation times, increased variability, and potentially higher radiation dose or image noise due to patient off-centering [19]. Automated patient positioning improves workflow efficiency and standardizes image acquisition [6]. Camera-based systems, especially RGB-D sensors, capture the external body surface and estimate patient pose, body contours, anatomical landmarks, table position, scan range, or patient isocenter [3,4,21]. These external cues are suitable for coarse positioning and body-region localization, but organ-specific scan planning requires knowledge of internal anatomy, which is not visible from the body surface.

Recent work has focused on estimating hidden anatomical information from external surface observations. Kats et al. investigated the localization of internal anatomical structures from depth images for automated scan planning [10]. Atici et al. studied the estimation of internal anatomy from body-surface point clouds, predicting organ-specific point representations from external surface geometry [1]. These works suggest that external body geometry can provide useful information about internal anatomy, but the predicted outputs remain tied to specific representations like organ landmarks or point sets.

In parallel, implicit neural representations (INRs) have become a powerful way to represent geometry as continuous coordinate-based functions. Occupancy-based models represent shape as a decision function predicting whether a spatial query point lies inside or outside the object [14,18]. Signed-distance-based models learn a continuous distance function with respect to the object surface [16]. SIREN introduced sinusoidal activation functions for INRs, enabling coordinate-based MLPs to represent fine spatial detail and derivatives of continuous signals [20]. Modulated periodic activations further extend this idea by conditioning periodic implicit networks on latent codes, allowing a shared network to represent families of signals rather than a single instance [13]. In medical imaging, INRs predict labels from spatial coordinates for memory-efficient and resolution-flexible segmentation. Implicit U-Net combines convolutional features with an implicit localization network for volumetric segmentation [12], while SCONet extends occupancy-based representations to multi-organ segmentation by estimating organ-wise occupancy in 3D space [9]. However, these methods typically condition on image- or contour-derived features. In contrast, the surface-to-viscera setting requires inferring internal anatomy from external observations alone.

We formulate internal anatomy estimation from body-surface point clouds as a surface-conditioned implicit reconstruction problem. Given an external body-surface point cloud, our model learns a conditional implicit representation of internal anatomy. It predicts anatomical class logits for arbitrary 3D query coordinates conditioned on a patient-specific latent code extracted from the body surface. This coordinate-based formulation enables resolution-flexible multi-organ reconstruction without tying the learned representation to a fixed output grid. The model combines a hierarchical Point Transformer encoder with a modulated SIREN decoder. We train and evaluate the model on a subset of 6,234 whole-body MRI scans from the German National Cohort (NAKO) study [2], using 4,780 subjects for training and 1,454 subjects for testing. The dataset provides



(a) Body-Surface Point Cloud (b) Query Points

Fig. 1. Data generation overview. (a) A synthetic top-down RGB-D observation is simulated from the MRI-derived body surface and converted into a body-surface point cloud. (b) Query points are sampled from automated multi-organ segmentation masks. We retain 20 consistently present internal anatomical structures and the surrounding background as target classes.

large-scale population-based whole-body imaging with substantial anatomical variability. Our experiments evaluate the reconstruction of 20 internal anatomical structures, including bones and soft-tissue organs, from body-surface point clouds alone.

2 Methods

2.1 Dataset Preparation

For each subject, the training data consists of an external body surface point cloud and labeled 3D query points sampled from reference segmentations. Data preparation involves two stages: body surface point cloud generation and query point sampling. An overview and description of these stages are shown in Fig. 1 and the following subsections.

Body Surface Point Cloud Generation Automated patient positioning systems use RGB-D cameras to capture the external body surface. In our dataset, we simulate this setting by assuming an RGB-D camera is positioned above the subject, oriented downwards towards the scanner table. The visible body surface is extracted from the 3D image grid and converted into surface coordinates. To obtain a fixed-size input, we apply farthest point sampling to generate a point cloud of size $(16{,}384 \times 3)$. The resulting body surface point cloud is transformed into a normalized coordinate system bounded by $[-1, 1]^3$.

Query Points Generation Since MRI scans are available for all subjects, automated 3D segmentation masks of anatomical reference labels are generated using TotalSegmentatorMRI [5], MRSegmentator [8], and TotalVibeSegmentator [7]. These combined masks cover 137 unique anatomical structures, including organs, blood vessels, bones, and tissue classes. We select a subset of 20 target labels based on consistent presence across all subjects and sufficient non-zero

Table 1. Curated set of 20 internal anatomical structures used in this study.

Soft Tissue Organs		**Bones**	
Right lung	Pancreas	Left hip	Sacrum
Left lung	Duodenum	Right hip	Left scapula
Liver	Aorta	Left clavicle	Right scapula
Right kidney	Heart	Right clavicle	L4 vertebra
Left kidney		Left femur	L5 vertebra
		Right femur	

boundary voxels. This results in a final set of 20 internal anatomical structures, listed in Table 1. The surrounding background is also retained as an additional target class.

Near-surface sampling is used because organ boundaries provide the most informative supervision for learning transitions between adjacent anatomical classes. For each structure, we use a base number of $N = 1{,}024$ samples: 90% are drawn from the organ surface and 10% are sampled uniformly from the organ's axis-aligned bounding box. Each sampled surface location is perturbed with zero-mean Gaussian noise at two scales, $\sigma \in \{1, 3\}$ voxels, resulting in two near-surface query points per surface sample. In addition, 4,096 background points are sampled uniformly from all background voxels. Each query coordinate is rounded to the nearest voxel and assigned the corresponding class label from the reference segmentation. All query points are transformed into the same normalized coordinate system as the body-surface point cloud.

2.2 Model Architecture

The proposed model learns a patient-specific implicit representation of internal anatomy conditioned on the external body surface. It consists of a hierarchical Point Transformer encoder that extracts a global latent representation from the body-surface point cloud and a SIREN-based decoder that predicts class logits for arbitrary 3D query coordinates. Fig. 2 shows an overview of the proposed architecture.

Encoder The encoder extracts a global shape representation from the body-surface point cloud to capture coarse anthropometric variability. Similar to the DeformingPointTransformer (DPT) [1], the point coordinates are processed by a hierarchical Point Transformer encoder [22]. The encoder has six hierarchical stages with local self-attention layers and residual connections. The first stage preserves the number of points, while the subsequent five stages downsample the point cloud by a factor of four and progressively increase the feature dimension to 512. Batch normalization is replaced by layer normalization within the Point Transformer layers and residual blocks, improving numerical stability. The final encoder stage produces 16 point-wise feature vectors with 512 features each. Unlike DPT, which retains these features for decoding, we apply global

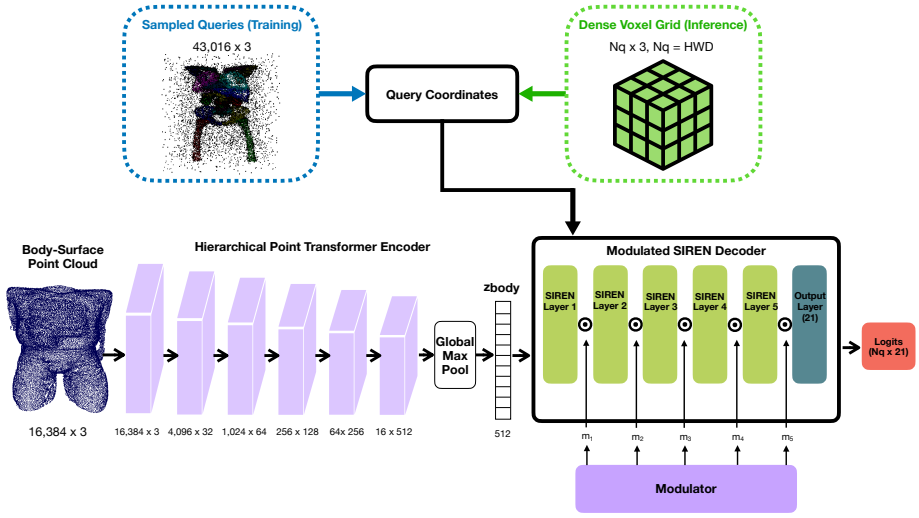


Fig. 2. Proposed model architecture overview. The hierarchical Point Transformer encoder processes the body-surface point cloud to obtain a global latent body code z_{body}. The modulated SIREN decoder predicts class logits for arbitrary spatial query coordinates conditioned on z_{body}. During training, $N_q = 43{,}016$ sampled query points are used per subject; their reference labels are shown only for visualization and serve as target labels for loss computation, not as model inputs. During inference, $N_q = HWD$ corresponds to all voxel coordinates of the dense target grid, which defines the queried output locations and is not a fixed model component. The symbol $\odot$ denotes element-wise modulation.

max pooling across the point dimension to aggregate them into a single 512-dimensional global shape representation z_{body}, a patient-specific descriptor of the complete body surface.

Decoder The decoder predicts class logits for each query point based on the global latent body code. The spatial coordinates of the query points are input to a SIREN [20], while z_{body} is processed by a feedforward modulator [13]. SIREN consists of five hidden layers with 256 features each. The first layer uses an initial angular frequency of $w_0 = 5$ to bias the decoder toward the lower-frequency spatial structure of coherent anatomical regions, rather than the fine-grained texture typically encountered in image reconstruction. The modulator generates one modulation vector for each hidden SIREN layer. After each layer, the hidden features are multiplied element-wise by the corresponding modulation vector. This implicitly conditions the function on the body shape of the individual subject. A final unmodulated output layer maps the hidden features to 21 class logits, representing background and 20 anatomical structures.

2.3 Training and Inference

During training, query points are sampled in 3D space using the strategy in Section 2.1, producing $N_q = 43{,}016$ query points per subject. The model predicts class logits for these points and is optimized using a combination of Dice loss [15] and cross-entropy loss.

During inference, the model predicts a dense segmentation map by querying the coordinates of all voxels in the target volume. The voxel grid defines the output query locations, not the learned representation. Thus, $N_q = HWD$ equals the volume's voxel count. Coordinates are evaluated in chunks to limit memory consumption, and each voxel is assigned to the class with the highest output logit.

3 Experiments

The model, implemented in PyTorch, was trained for 300 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of eight on an NVIDIA H100 GPU with 80 GB of memory. Using anonymized NAKO data under the applicable data-use agreement, we trained on 4,780 subjects and evaluated on an independent test set of 1,454 subjects. Each subject was represented by a body-surface point cloud and a corresponding volumetric reference segmentation. All volumes had dimensions of $320 \times 260 \times 316$ voxels ($X \times Y \times Z$) with an isotropic voxel spacing of 1 mm.

For comparison, we included two learned baselines from DPT [1]: the point-set-based DPT and its volumetric convolutional autoencoder (PointCAE) baseline, as well as the non-learned mean-shape baseline. PointCAE provides a conventional dense volumetric reference, reconstructing internal anatomy using a 3D convolutional encoder-decoder from a rasterized body-surface input. The mean-shape baseline uses the population-average anatomical template as the same prediction for every test subject, providing a reference for quantifying the benefit of patient-specific conditioning.

Volumetric predictions were evaluated using the average symmetric surface distance (ASSD) and the 95th-percentile Hausdorff distance (HD95), both reported in millimetres. ASSD is conceptually related to the Chamfer Distance (CD) reported by DPT, but is computed on surfaces extracted from volumetric segmentations rather than on sampled point sets. This allows evaluation of the entire reconstructed anatomical surface, while DPT is evaluated on its native point-set representation.

We additionally investigated zero-shot inference on ten real depth observations acquired with an RGB-D system. The depth maps were converted into body-surface point clouds using back-projection with calibrated camera intrinsics, subtraction of corresponding empty-table acquisitions, and cropping guided by available 2D body landmarks to match the anatomical field of view of the synthetic inputs. The resulting point clouds were subsequently sampled and normalized as described in Section 2.1. The frozen model was applied without retraining or fine-tuning. Since no corresponding volumetric reference annotations

were available, this experiment provides a qualitative feasibility demonstration rather than a quantitative evaluation of anatomical accuracy.

4 Results

Results of the quantitative evaluation are reported in Tables 2 and 3. The mean-shape baseline yields the largest surface errors, highlighting the limitations of using the same population-average anatomy for all subjects. Conditioning on the individual body surface substantially improves the reconstruction: PointCAE achieves a mean CD of 8.07 mm, while DPT further reduces it to 4.78 mm. Our method achieves a mean ASSD of 2.95 mm. For HD95, which is reported for all four methods, the same trend is observed. The error decreases from 31.07 mm for the mean shape to 24.25 mm for PointCAE and 12.06 mm for DPT, while our approach further reduces it to 9.26 mm.

The methods differ in their output representation: DPT predicts a fixed-size point set of 4,096 points per structure, whereas our model defines a coordinate-based implicit function that is evaluated on the full target volume to obtain dense segmentations. This distinction should be considered when interpreting surface-based metrics, particularly HD95, which is sensitive to sparse predictions and localized outlier surface points. The per-structure comparison in Table 3 shows that our method yields lower CD/ASSD values for all 20 anatomical structures and lower HD95 values for 17 of them. Notable improvements are observed for the aorta (19.42 mm to 8.01 mm) and for large soft-tissue organs such as the lungs and liver.

Fig. 3 shows a representative test case selected as the subject closest to the median aggregate ASSD. The qualitative result illustrates that the model captures the overall anatomical configuration from the body-surface input. The surface distance map shows mostly low errors, with localized deviations around abdominal structures and the pelvic region, indicating that fine surface details and organ interfaces remain challenging.

Post-hoc visual inspection identified erroneous heart reference annotations in 22 of the 1,454 test subjects. The primary results reported above include all test subjects. To assess the influence of these annotation errors, we performed a sensitivity analysis for the heart class by excluding the affected cases, resulting in 1,432 valid heart annotations. After exclusion, the heart ASSD decreased from 5.07 mm to 3.87 mm, and HD95 decreased from 18.48 mm to 17.02 mm. This indicates that erroneous reference annotations negatively affected the reported heart metrics.

A complete dense segmentation volume was generated in approximately 6.18 s per subject. Encoding the body-surface point cloud accounted for only 0.05 s, while the remaining runtime was primarily associated with evaluating the implicit decoder over the dense query grid.

Zero-shot inference produced volumetric predictions for all ten real depth observations. Fig. 4 shows the cases with the smallest and largest external torso extent, selected based on the input observations before inspecting the predic-

Table 2. Aggregate structure-level performance. Baseline values are from [1]. Values are mean ± standard deviation across 20 structure-wise means. DPT reports CD, whereas our method reports ASSD; HD95 is reported for both. Lower is better.

Metric	Mean Shape	PointCAE	DPT	Ours
CD/ASSD (mm)	11.20 ± 1.26	8.07 ± 1.01	4.78 ± 1.33	$\mathbf{2.95 \pm 0.98}$
HD95 (mm)	31.07 ± 4.24	24.25 ± 3.88	12.06 ± 3.64	$\mathbf{9.26 \pm 3.24}$

Table 3. Per-structure comparison. DPT values are from [1]. DPT uses CD, whereas our method uses ASSD; HD95 is reported for both. Values are in millimetres, bilateral structures are averaged, and bold indicates lower error.

Structure	DPT CD	Ours ASSD	DPT HD95	Ours HD95
Aorta	6.30	**2.78**	19.42	**8.01**
Clavicles (L/R)	2.52	**1.64**	**5.85**	6.06
Duodenum	4.61	**3.47**	12.29	**9.59**
Femurs (L/R)	4.57	**2.70**	10.81	**8.54**
Heart	7.75	**5.07**	**17.00**	18.48
Hips (L/R)	4.49	**2.46**	10.15	**7.47**
Kidneys (L/R)	4.91	**4.42**	14.05	**13.33**
Liver	6.74	**4.26**	16.60	**12.73**
Lungs (L/R)	6.06	**2.86**	14.66	**8.17**
Pancreas	4.88	**3.72**	**11.97**	12.87
Sacrum	5.01	**2.61**	13.61	**8.09**
Scapulae (L/R)	3.36	**1.86**	7.73	**7.05**
Vertebrae L4–L5	4.22	**2.63**	11.92	**7.06**

tions. The case with the larger torso extent exhibits a visibly broader predicted anatomical configuration, particularly in the transverse extent of the abdominal organs and pelvis. In contrast, the prediction for the smaller case appears more compact.

5 Discussion and Conclusion

The proposed model demonstrates that internal anatomy can be reconstructed from external body-surface geometry using a conditional implicit representation. In contrast to fixed point-set predictions, the coordinate-based decoder can be queried at arbitrary spatial locations and produces dense volumetric segmentations. The results indicate that a compact global body code, obtained by pooling hierarchical Point Transformer features, retains sufficient patient-specific information to condition the modulated SIREN decoder.

The method achieved lower HD95 values for 17 of the 20 anatomical structures, and the ASSD values of our volumetric predictions were numerically lower than the CD values reported for DPT across all structures. The qualitative zero-shot experiment further demonstrates the technical feasibility of applying the complete pipeline to body-surface point clouds extracted from real depth observations without retraining or fine-tuning. Moreover, the predicted anatomical

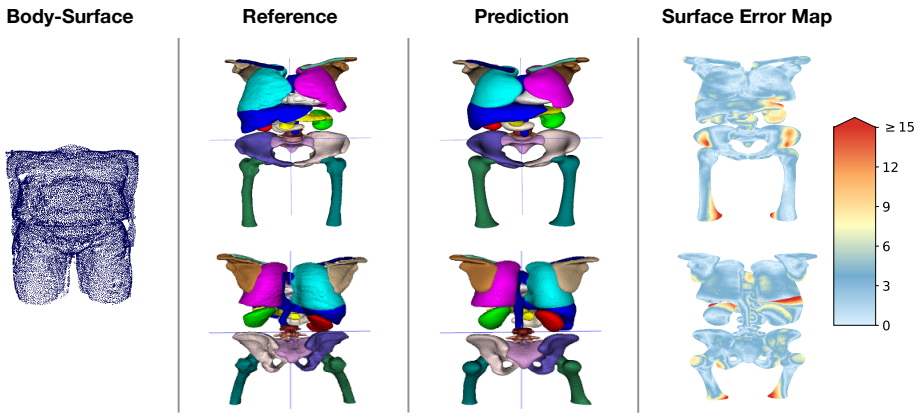


Fig. 3. Qualitative reconstruction of a representative test subject. Columns show body-surface input, reference anatomy, prediction, and prediction-to-reference surface-distance map. Upper and lower rows show anterior and posterior views. Distances are shown in millimetres and clipped at 15 mm. The case is closest to the median aggregate ASSD.

configurations varied with the observed external torso geometry. Since no corresponding volumetric reference annotations were available, this experiment does not establish anatomical accuracy. Nevertheless, it suggests the feasibility of using real depth observations for future anatomy-aware clinical workflow support.

Future work should evaluate the method in larger real-world cohorts using paired depth observations and volumetric reference annotations. Robustness to sensor noise, occlusions and incomplete surface coverage should also be investigated systematically. As a separate methodological direction, signed distance functions could be investigated as an implicit output representation, as their explicit encoding of distance to anatomical boundaries may provide stronger geometric supervision.

In conclusion, the proposed surface-conditioned implicit model enables dense, patient-specific reconstruction of internal anatomy from external body-surface geometry alone, demonstrating its potential as an anatomy-aware component of automated patient positioning workflows.

Acknowledgments. We gratefully acknowledge the financial support by German Research Foundation: DFG, HE 7364/10-1, project number 500498869.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

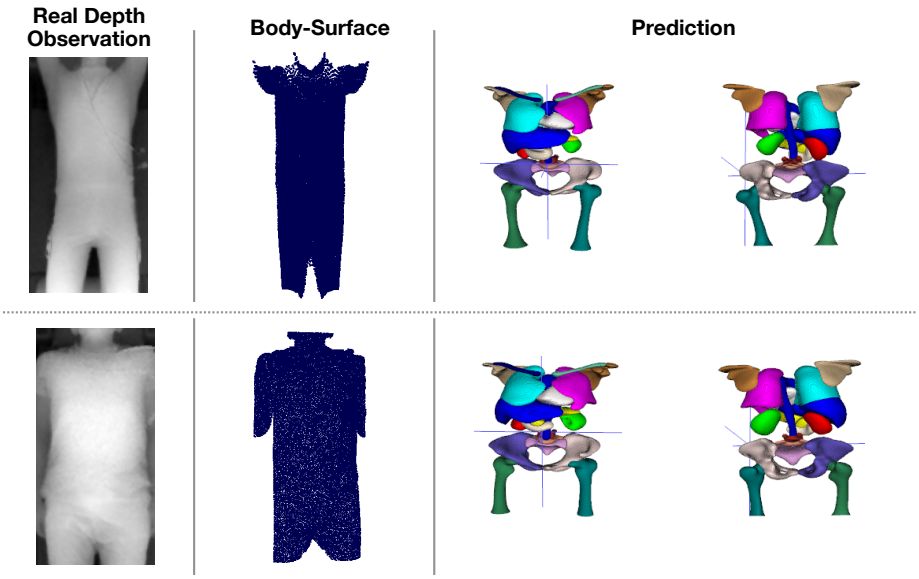


Fig. 4. Qualitative zero-shot inference on real depth observations. The upper and lower rows show the cases with the smallest and largest external torso extent, respectively. From left to right, the columns show the cropped real depth observation, the extracted body-surface point cloud and the predicted internal anatomy from anterior and posterior views. All predictions are displayed using identical camera settings and spatial scaling.

References

1. Atici, S.F., Kats, E., Mensing, D., Heinrich, M.P.: From surface to viscera: 3d estimation of internal anatomy from body surface point clouds. In: Medical Imaging with Deep Learning (2026)
2. Bamberg, F.: Whole-body magnetic resonance imaging in the german national cohort (nako): Design & current status. European Journal of Public Health **32** (2022)
3. Booij, R., Budde, R.P., Dijkshoorn, M.L., van Straten, M.: Accuracy of automated patient positioning in ct using a 3d camera for body contour detection. European Radiology **29**(4), 2079–2088 (2019)
4. Booij, R., van Straten, M., Wimmer, A., Budde, R.P.: Automated patient positioning in ct using a 3d camera for body contour detection: accuracy in pediatric patients. European radiology **31**(1), 131–138 (2021)
5. D'Antonoli, T.A., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiß, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., et al.: Totalsegmentator mri: robust sequence-independent segmentation of multiple anatomic structures in mri. arXiv preprint arXiv:2405.19492 (2024)
6. Gang, Y., Chen, X., Li, H., Wang, H., Li, J., Guo, Y., Zeng, J., Hu, Q., Hu, J., Xu, H.: A comparison between manual and artificial intelligence–based automatic positioning in ct imaging for covid-19 patients. European Radiology **31**(8), 6049–6058 (2021)

7. Graf, R., Platzek, P.S., Olga Riedel, E., Ramschütz, C., Starck, S., Möller, H.K., Atad, M., Völzke, H., Bülow, R., Schmidt, C.O., et al.: Totalvibesegmentator: Full body mri segmentation for the nako and uk biobank. arXiv e-prints pp. arXiv–2406 (2024)
8. Häntze, H., Xu, L., Mertens, C.J., Dorfner, F.J., Donle, L., Busch, F., Kader, A., Ziegelmayer, S., Bayerl, N., Navab, N., et al.: Mrsegmentator: Multi-modality segmentation of 40 classes in mri and ct. arXiv preprint arXiv:2405.06463 (2024)
9. Jouvencel, M., Kéchichian, R., Digne, J., Valette, S.: Sconet: Convolutional occupancy networks for multi-organ segmentation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
10. Kats, E., Geißler, K., Hirsch, J.G., Heldman, S., Heinrich, M.P.: Internal organ localization using depth images: A framework for automated mri patient positioning. In: BVM Workshop. pp. 324–329. Springer (2025)
11. Koken, P., Dries, S.P., Keupp, J., Bystrov, D., Pekar, V., Börnert, P.: Towards automatic patient positioning and scan planning using continuously moving table mr imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **62**(4), 1067–1072 (2009)
12. Marimont, S.N., Tarroni, G.: Implicit u-net for volumetric medical image segmentation. In: Annual Conference on Medical Image Understanding and Analysis. pp. 387–397. Springer (2022)
13. Mehta, I., Gharbi, M., Barnes, C., Shechtman, E., Ramamoorthi, R., Chandraker, M.: Modulated periodic activations for generalizable local functional representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14214–14223 (2021)
14. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
16. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
17. Pekar, V., Bystrov, D., Heese, H.S., Dries, S.P., Schmidt, S., Grewer, R., Den Harder, C.J., Bergmans, R.C., Simonetti, A.W., Van Muiswinkel, A.M.: Automated planning of scan geometries in spine mri scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 601–608. Springer (2007)
18. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: european conference on computer vision. pp. 523–540. Springer (2020)
19. Saltybaeva, N., Schmidt, B., Wimmer, A., Flohr, T., Alkadhi, H.: Precise and automatic patient positioning in computed tomography: avatar modeling of the patient surface using a 3-dimensional camera. Investigative Radiology **53**(11), 641–646 (2018)
20. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in neural information processing systems **33**, 7462–7473 (2020)

21. Teixeira, B., Singh, V., Tamersoy, B., Prokein, A., Kapoor, A.: Automated ct lung cancer screening workflow using 3d camera. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–431. Springer (2023)
22. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021)