Dose robustness of deep learning models for anatomic segmentation of CT images

Artyom Tsanda [a,b,*], Hannes Nickisch[b], Tobias Wissel[b], Tobias Klinder[b], Tobias Knopp[a,c], and Michael Grass[b]

[a]Hamburg University of Technology, Institute for Biomedical Imaging, Hamburg, 21073, Germany
[b]Philips Innovative Technologies, Hamburg, 22335, Germany
[c]University Medical Center Hamburg-Eppendorf, Section for Biomedical Imaging, Hamburg, 22529, Germany
[*]Corresponding author: Artyom Tsanda, artyom.tsanda@tuhh.de

Abstract

Purpose: The trend towards lower radiation doses and advances in CT reconstruction may impair operation of pre-trained segmentation models, thereby giving rise to the problem of estimating dose robustness of existing segmentation models. Previous studies addressing the issue suffer either from a lack of registered low- and full-dose CT images or from simplified simulations.

Approach: In this work, we employ raw data from full-dose acquisitions to simulate low-dose CT scans, avoiding the need to rescan a patient. The accuracy of the simulation is validated using a real CT scan of a phantom. We consider down to 20% reduction of radiation dose, for which we measure deviations of several pre-trained segmentation models from the full-dose prediction. Additionally, compatibility with existing denoising methods is considered.

Results: The results reveal surprising robustness of the TotalSegmentator approach, showing minimal differences at the pixel level even without denoising. Less robust models show good compatibility with the denoising methods, which help to improve robustness in almost all cases. With the CNN-based denoising, the median Dice between low-and full-dose data does not fall below 0.9 (12 for the Hausdorff distance) for all but one model. We observe volatile results for labels with effective radii less than 19 mm and improved results for contrasted CT acquisitions.

Conclusion: The proposed approach facilitates clinically relevant analysis of dose robustness for human organ segmentation models. The results outline robustness properties of a diverse set of models. Further studies are needed to identify robustness of approaches for lesion segmentation and to rank the factors contributing to dose robustness.

Keywords— Low-dose CT, Semantic segmentation, Denoising, Deep learning

1 Introduction

Computed tomography (CT) is an important medical imaging modality widely adopted in many clinical applications. It is commonly used for detecting injuries, tumors, infections, and abnormalities. Since the analysis of CT images is time consuming, automated CT annotation and quantification remains an active area of research.

Semantic image segmentation is a significant step in CT analysis and often a prerequisite for subsequent quantitative assessments and clinical decision making [11]. For this reason, great efforts are being made to alleviate the burden of manual segmentation from radiologists. In recent years, the performance of algorithms for automated CT segmentation has considerably increased due to the adoption of deep neural networks (also deep learning models) [30]. Neural networks implicitly derive useful segmentation features from pairs of CT images and the corresponding ground truth segmentations through the training process. Therefore, the training data need to be representative with respect to the following application.

At the same time, CT acquisitions expose a patient to X-ray radiation, which may have irreversible consequences to the human body. For this reason, usage and development of CT systems is guided by the ALARA ("as low as reasonably achievable") principle [22]. This implies the inclination to lower tube currents, thus lower dose, whenever possible. Dose reduction mainly leads to a decreased signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR), limiting the diagnostic quality of the resulting images. Therefore, the research community actively investigates CT denoising algorithms to increase image quality of low-dose CT (LDCT) scans [8].

To effectively increase the SNR in LDCT scans, several methods have been developed in recent years. They can be categorized into projection [32, 39] and image denoising [29, 4, 41]. Solving the denoising problem for

raw projections is advantageous as the noise model is known. However, in the projection space, denoising is applied on top of the Radon transform, where small image details are hidden in the line integrals. As a result, images reconstructed from denoised projections run the risk of being blurred. On the other hand, denoising reconstructed images reveals opportunities for formulating an optimization problem in terms of the desired image quality, but leads to an unknown noise distribution.

Although there are approaches considering semantic segmentation and LDCT image denoising jointly [10, 21], they are usually developed independently and may introduce discrepancies between expected and actual image characteristics. Therefore, the risk of incompatibility exists. Retrospective data used to train segmentation models may contain neither new reconstruction options nor lower dose levels. This can lead to unexpected predictions of a segmentation model when applied to low-dose data.

We propose a method to assess robustness of existing semantic segmentation models against dose reduction in CT images. Our approach relies on low-dose simulations conducted on raw data, which allows reliable estimations per patient without the need for additional rescanning. This paper has two major contributions:

- We provide quantitative estimations of dose robustness for existing deep neural networks trained for semantic segmentation of human organs in CT scans. These estimations can help assess generalization and identify limitations when reducing the dose in CT scans without retraining existing segmentation models.

- We quantitatively investigate the impact of denoising methods applied to low-dose CT scans on segmentation results.

Application areas of our research include radiation therapy planning using low-dose CT protocols [35], segmentation of cancer screening data [6], segmentation of CT data for interventional therapy planning [26], and analysis of novel LDCT protocols using existing deep learning models [1].

This paper is an extension of the conference abstract [38]. Compared with the abstract, we made the following major extensions. Firstly, we significantly extended the set of segmentation models included in the study. Secondly, the effect of denoising methods on dose robustness was investigated. Thirdly, more ablation studies were added.

2 Related Work

Robustness of deep neural networks (DNNs) has received increasing attention in recent years due to their wide adoption in real-world applications. Although DNNs achieve state-of-the-art performance in various tasks, their black-box nature poses significant challenges for understanding their limitations. A data-driven approach to produce DNNs and verify their performance may lead to poor generalization, e.g., in the case of overfitting, and even malicious attacks [12]. Therefore, many studies are aimed to investigate and improve robustness of DNNs with respect to realistic input perturbations, enabling them to operate safely and reliably in diverse environments.

Due to the high availability of massive data volumes and realistic corruption models, the problem has been more extensively researched in the domain of natural images. Hendrycks *et al.* [18] enriched the validation set of ImageNet [7] with diverse corruptions (Gaussian noise, impulse noise, frosted glass blur etc.) of varying intensities (the ImageNet-C dataset). They showed that advances in performance from AlexNet [25] to ResNet [16] are not translated into robustness at the same scale. Kamann *et al.* [24] explored robustness of semantic segmentation models. They extended the set of corruptions used in ImageNet-C with noise, blur and geometric distortions specific to the camera. The authors considered the DeepLabv3+ architecture [5] with various backbones. In contrast to the previously mentioned work, the results revealed a correlation between model performance and its robustness. They also highlighted the substantial impact of noise corruption on model performance.

Contrary to natural images, in the domain of medical images, specifically in CT, robustness studies are facing difficulties with both data availability and realistic corruptions. During the 2016 Low Dose CT Grand Challenge [33] low-dose CT images were simulated using projection data and used to assess the impact of denoising algorithms on the ability of radiologists to identify lesions inside the liver. For the top-performing denoising methods, observer performance was comparable to that of the full-dose setting. Hammond *et al.* [14] acquired CT scans of a male swine using various low-dose protocols. The acquired lung volumes were segmented using an intensity-based segmentation algorithm. They reported visually successful segmentation results, and for low-dose scans, the selected quantitative parenchymal and airway measurements remained relevant to pulmonary disease characterization. Hooper *et al.* [19] investigated robustness of a 3D classification network (121-layer DenseNet [20]) trained to triage head CT data. The authors reprojected available CT volumes in axial geometry and simulated reduced tube current, limited angle and sparse view artifacts at different scales (4x, 8x, 16x). The first was done by adding Gaussian white noise with the adjusted variance to the projections. Although the

simulation of acquired projections may introduce additional errors, and the physics of CT acquisition implies the Poisson noise model, the authors found the performance of the model on all levels of tube current reduction to be surprisingly stable. Liu *et al.* [31] included adversarial noise realizations as well as adversarial synthetic lung nodules in the training of a nodules detection network. They showed that the resulting model also became more robust against uniform and Poisson noise models applied in the image space. Aiello *et al.* [1] investigated the applicability of existing deep learning models for lungs and COVID-19 lesions segmentation trained on full-dose CT scans to low-dose scenarios. Due to the lack of registered pairs of low- and full-dose data, the authors compared the statistics calculated based on the resulting segmentation masks, specifically the COVID-19 volume percentage. The results showed a high level of agreement between low- and full-dose predictions.

In all of the aforementioned studies, we observe a trade-off between the use of actual and simulated low-dose data. The use of actual CT data imposes additional limitations on the dataset size, often requiring an extra registration step. On the other hand, simulations allow scaling experiments, albeit at the cost of using simplified simulations due to the absence of raw data. Consequently, the applicability of the results to real-life scenarios becomes a matter of concern.

3 Materials and Methods

Our method relies on simulations of low-dose CT acquisitions. For this purpose, we use a dataset with raw projections, which includes tube currents and reference photon counts. We reconstruct the data using several denoising options and subsequently apply pre-trained segmentation models to these reconstructions. As the data is not annotated, we define robustness as a measure of deviation from the full dose prediction, and calculate it using segmentation metrics. In this section, we introduce each step of this pipeline.

3.1 Low Dose Simulation

To avoid potential rescanning of a patient, we simulate low-dose CT acquisitions from full-dose projections. Since the noise model in the projection space is known, we can achieve accurate simulation results without additional radiation exposure [34].

Having an object with attenuation μ and the mean number of photons $\overline{n}_0^\alpha$ emitted at current α, according to Beer's law, the mean number of photons hitting the detector element $\overline{n}^\alpha$ along ray $r(i)$ is given by[3]

$$\overline{n}^\alpha = \overline{n}_0^\alpha \cdot \exp\left(-\int_{r(i)} \mu \, dr\right), \tag{1}$$

where the integral $\int_{r(i)} \mu \, dr$ is referred to as the line integral l^α and latter used in the reconstruction algorithms.

Due to the nature of the radiation process and photon-matter interaction, both the number of photons at the emitter and the receiver obey Poisson statistics. Electronics inside the detector element also introduce noise to the measurements, which can be well-described using the Gaussian distribution; however, in our experiments, we neglect it. As a result, the acquired measurements follow the distribution

$$n^\alpha \sim \mathcal{P}(\overline{n}^\alpha). \tag{2}$$

With known mean values, one can simulate any tube current, but in reality, only noise realizations are available. To estimate the number of photons at a lower current β, we follow the approach proposed by Žabić *et al.* [42] and sample new photon counts $n^{\alpha\to\beta}$ from the following distribution:

$$n^{\alpha\to\beta} \sim \frac{\alpha-\beta}{\alpha}\mathcal{P}\left(\frac{\beta}{\alpha-\beta}n^\alpha\right). \tag{3}$$

Since the initial tube current cannot be changed and is dictated by the already acquired CT data, we will operate in terms of relative dose level defined as

$$\text{dose level} = \frac{\beta}{\alpha} * 100\%. \tag{4}$$

In this study, the input data represent line integrals l^α measured for each detector element and each detector position for the current α. Using Equation (1), we convert the line integrals into the photon counts n^α. The number of emitted photons $\overline{n}_0^\alpha$ is measured using air scans. Following Equation (3), we sample new photon counts n^β for a lower current β. With $\overline{n}_0^\beta = \overline{n}_0^\alpha \cdot \beta/\alpha$, the resulting photon counts are converted back to line integrals l^β and used for the reconstruction.

The accuracy of the considered low-dose model was confirmed experimentally using axial phantom CT scans. The phantom was scanned several times at different tube currents. Between acquisitions, the phantom was not

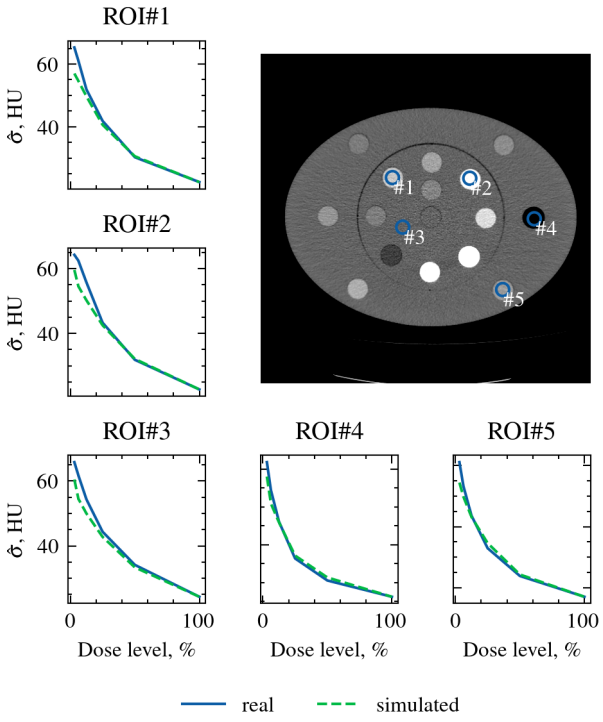


Figure 1: A full-dose CT scan of the phantom used to validate the low-dose simulation along with the selected ROIs. The sampled standard deviation $\hat{\sigma}$ is calculated for each ROI and for each dose level. The deviation across dose levels for each ROI is shown next to the image.

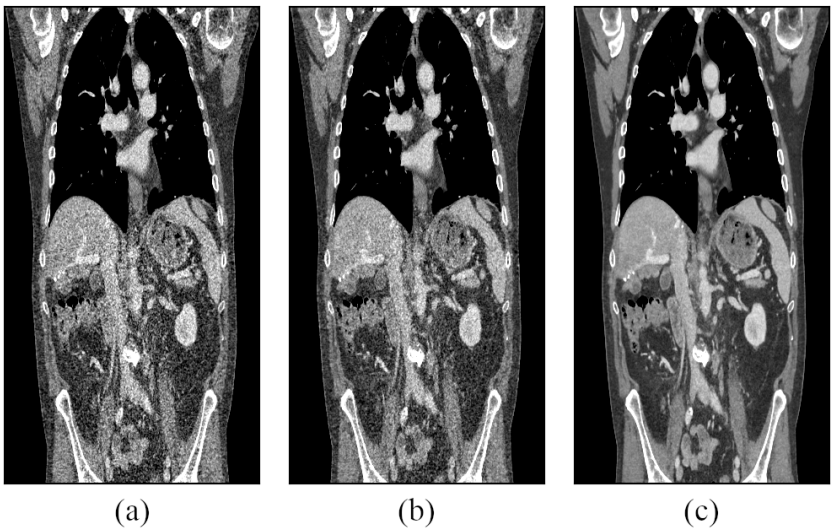


Figure 2: Examples of reconstructed images at 20% dose level with different denoising methods applied: (a) without, (b) iterative and (c) CNN. Level/Window is set to 50/500.

moved; thus, the resulting CT images were registered. Given projections corresponding to the highest tube current, we simulated low dose scans with the corresponding tube currents. To compare the results, we selected several regions of interest (ROIs), as shown in Fig. 1. Within the ROIs, we calculated the standard deviations for each current. In this study, we limit the range of dose levels with 20% so that the average relative error between sampled standard deviations calculated for simulated and real scans within each ROI does not exceed 1%.

3.2 Semantic segmentation

Our study includes several existing deep learning models trained on various CT datasets for 3D semantic segmentation of human organs.

The majority of the selected segmentation models are based on the nnU-Net framework[23]. nnU-Net aims to automate the configuration (i.e., pre-processing, network architecture, training and post-processing) of deep learning-based segmentation. For instance, the framework automatically resamples and normalizes data based on the statistics calculated across all training cases. During training, nnU-Net applies the following augmentations randomly: rotations, scaling, Gaussian noise, Gaussian blur, change of brightness and contrast, simulation of low resolution, gamma correction, and mirroring. Among these, the Gaussian noise augmentation may be the most relevant for dose robustness. It is applied to normalized data with a 15% probability, and variance is drawn from $U(0, 1)$. Recently, a large dataset of CT scans for organ segmentation has been released[40]. When combined with nnU-net, it led to a solution called TotalSegmentator[40]. TotalSegmentator consists of five segmentation models for different organ groups. We analyze them individually as well as in combination.

In addition, we include two recently proposed transformer-based[9] methods: UNETR[15] and Swin UNETR[37]. The latter achieves state-of-the-art results on several segmentation datasets, including Medical Segmentation Decathlon[36]. Both approaches use random flips, rotations and intensities shifting during training. We consider 3 Swin UNETR models with different number of parameters. The largest model *swin-unetr-base* is initialized with the weights after self-supervised pre-training, other two models are trained from scratch.

The considered models mostly employ contrasted CT scans at portal venous phase for training. Regarding the rest, KiTS2021[17] contains contrasted CT scans at the late arterial phase, SegTHOR[27] contains scans with or without intravenous contrast, and TotalSegmentator[40] includes a variety of contrast phases as well as non-contrasted CT images.

The taxonomy of the models included in this study is shown in Table 1. The pre-trained models and evaluation scripts are obtained from the corresponding official repositories [1,2,3]. No fine-tuning on low-dose data is performed. In the case of nnU-Net, only 3D models with the highest resolution are used for evaluation. If multiple models are available for different folds of cross-validation, only the first model is selected. Because the scope of the segmentation is limited to human organs, no other predicted labels are included. For instance, *nnunet-kits* predicts labels not only for the kidney, but also for tumors. However, in this case, only the kidney segmentation is considered, while all other labels are ignored.

3.3 CT denoising

Noise is an essential attribute of CT images. Even without lowering the dose, existing acquisition protocols yield images corrupted by noise. For this reason, denoising algorithms have been a research topic for many years to assist clinicians in better resolving low-contrast details.

In this study, iterative and CNN-based denoising methods will be a part of the evaluation implemented within the corresponding iDose^4[2] and Precise Image[13] reconstructions (Philips Healthcare, Cleveland, Ohio, USA). iDose^4 includes noise reduction in both the projection and image space. Addressing noise in raw projections copes better with streak and bias artifacts. The following image-based denoising preserves the noise power spectrum and underlying edges. The CNN-based denoising method represents a network trained in a supervised manner with simulated low-dose CT data. Reconstructions without denoising in the image space are also included in this study.

3.4 Data

This study was performed in compliance with the local Institutional Review Board of Tel Aviv Sourasky Medical Center, Israel. Informed consent was obtained from all subjects and/or their legal guardian(s). All the methods were performed in accordance with the relevant guidelines and regulations. All scans were performed on the Philips Spectral CT 7500 System as part of the regulatory approval pathway required by governmental regulatory bodies.

[1]https://github.com/MIC-DKFZ/nnUNet/tree/nnunetv1
[2]https://github.com/Project-MONAI/research-contributions
[3]https://github.com/wasserth/TotalSegmentator

Table 1: Taxonomy of segmentation models included into the study.

Model Name	Training Dataset	#training-scans	Architecture	#parameters	#classes
totsegm-all	TotalSegmentator [40]	consists of the five models below			
totsegm-organs	TotalSegmentator [40]	1082	U-Net	31M	18
totsegm-vertebrae	TotalSegmentator [40]	1082	U-Net	31M	25
totsegm-cardiac	TotalSegmentator [40]	1082	U-Net	31M	19
totsegm-muscles	TotalSegmentator [40]	1082	U-Net	31M	22
totsegm-ribs	TotalSegmentator [40]	1082	U-Net	31M	25
nnunet-liver	MSD [36]	104	U-Net	31M	3
nnunet-abdmn	BTCV [28]	24	U-Net	31M	14
nnunet-pancreas	MSD [36]	224	U-Net	31M	3
nnunet-spleen	MSD [36]	32	U-Net	31M	2
nnunet-thor	SegTHOR [27]	32	U-Net	31M	5
nnunet-kits	KiTS2021 [17]	240	U-Net	31M	4
swin-unetr-base	BTCV [28]	24	Swin UNETR	62M	14
swin-unetr-small	BTCV [28]	24	Swin UNETR	16M	14
swin-unetr-tiny	BTCV [28]	24	Swin UNETR	4M	14
unetr	BTCV [28]	24	UNETR	93M	14

The data represent raw projections of the abdominal area in helical geometry with 8 cm collimation and pitch factor 1.38. The reconstruction is done using 420 mm field of view (FOV) with pixel numbers along X and Y set to 512. The slice thickness is 1 mm with the slice increment 0.5 mm. Approximately half of the CT studies include contrast in variable phases injected either orally or intravenously. Data of 42 patients with a total of 99 scans are used for the evaluation in this study. The dataset contains both scans of healthy patients and those with pathologies.

We consider 5 dose levels: 20%, 40%, 60%, 80%, 100%. Low-dose simulation and CT reconstruction is performed for each dose factor and each denoising method (w/o denoising, iterative, CNN). The resulting dataset consists of 1485 CT scans. Although the data come from a spectral scanner, we only consider conventional images. The reason for the simplification is the following segmentation operating only on conventional images. Examples of the reconstructed images are shown in Fig. 2. The reconstructed data are passed as inputs to the segmentation models.

3.5 Evaluation

Since reference annotations are not available, dose robustness is defined as the deviation of low-dose segmentation results from the full-dose results. It is measured by calculating metrics between segmentations for a reduced and 100% dose level.

In this paper, we employ the Dice similarity coefficient (Dice) and Hausdorff Distance 95% (HD95) to assess the difference between two segmentation results. HD95 represents the 95^{th} percentile of surface distances between two segmentation masks. The metrics are calculated as follows:

$$\text{Dice}(x_i, y_i) = \frac{2\sum_i x_i * y_i}{\sum_i x_i + \sum_i y_i}, \tag{5}$$

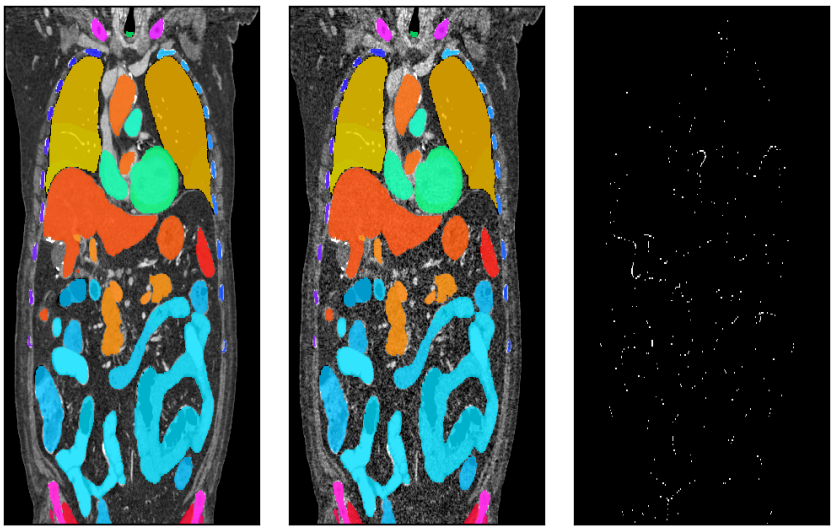
$$\text{HD95}(x_i, y_i) = P_{95}(d(x_i, y_j), d(y_j, x_i)), \tag{6}$$

where x_i and y_j denote voxels in 3D binary segmentation masks. In the case of multiple labels, we average the values produced for each one of them.

4 Results

In Table 2, we compare robustness of the segmentation models when reducing dose level to 20%. The metrics are grouped according to the denoising options. The median and the median absolute deviation were used to aggregate the results as some metrics fall into the extremes of the value range. The *swin-unetr-base*, *nnunet-lung*, *swin-unetr-tiny* models are the least robust, whereas the models from TotalSegmentator are the most robust. Advances in denoising methods help improve robustness in almost all cases, and in some cases, it plays

a decisive role. For example, for the *swin-unetr-base* model, the median Dice improves from 0 for scans without denoising to 0.8 for scans with CNN-based denoising. With the CNN-based denoising, the median Dice between low- and full-dose data does not go below 0.9 (12 for the Hausdorff distance) for all but one model. An example of segmentation results calculated for 20% and 100% dose level images for the *totalsegm-all* model is shown in Fig. 3. Examples for other models can be found in supplementary Figures S1-S11.



(a) (b) (c)

Figure 3: A segmentation example for *totalsegm-all* calculated for (b) 20% and (a) 100% dose level images with iterative denoising applied. The last image (c) shows the difference between the two segmentations.

5 Discussion

5.1 Robustness of the TotalSegmentator approach

As shown in the results, the TotalSegmentator model outperforms other approaches in terms of robustness. Even without additional denoising techniques, it reproduces the segmentation results at lower doses.

The key feature of this approach is the dataset size and diversity, which is four times larger than the next largest dataset, as shown in Table 1. In addition, the training data contain labeled low-dose CT scans. To demonstrate this, we calculate the standard deviation within the liver region. Assuming a constant CT value of the liver, the statistics allows to assess the dose. Figure 4 shows that the distribution of the standard deviations corresponding to the TotalSegmentator data has a heavy tail toward higher noise levels. The distribution partially overlaps with the one for 20% dose level. Another key attribute of TotalSegmentator is the large number of predicted classes, which may provide better reinforcement during training, leading to a more robust latent space. Finally, as the model is based on the nnU-Net framework, it was trained using Gaussian noise augmentation. However, other nnU-Net models, such as *nnunet-liver* or *nnunet-abdmn*, do not achieve similar results making the augmentation not the sole contributor to dose robustness. We consider the aforementioned features of TotalSegmentator as potential root causes of its robustness, however, isolated experiments are required to confirm that.

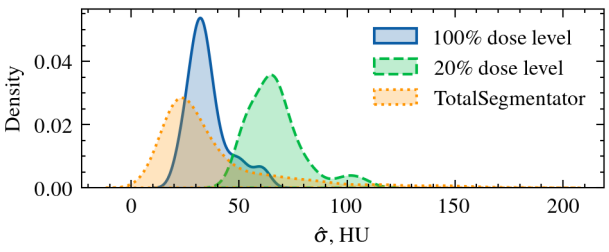


Figure 4: Kernel density estimations of sampled standard deviations calculated inside the liver region for each CT scan from different datasets: 20% and 100% dose data reconstructed using iterative denoising and TotalSegmentator training data.

Table 2: Differences between segmentation masks calculated for the corresponding 20% and 100% dose level CT scans. The difference is measured with the Dice score and the Hausdorff distance (95 percentile). The median and the median absolute deviation is calculated across all CT images for a given pair of a model and denoising method.

Metric	Model Name	w/o denoising	iterative	CNN
Dice	totsegm-vertebrae	0.988 ± 0.002	0.988 ± 0.002	0.992 ± 0.002
	totsegm-organs	0.986 ± 0.012	0.987 ± 0.011	0.990 ± 0.009
	totsegm-muscles	0.986 ± 0.004	0.986 ± 0.004	0.990 ± 0.005
	totsegm-all	0.982 ± 0.009	0.983 ± 0.008	0.988 ± 0.006
	nnunet-spleen	0.975 ± 0.012	0.980 ± 0.010	0.988 ± 0.007
	totsegm-cardiac	0.981 ± 0.010	0.982 ± 0.010	0.986 ± 0.008
	totsegm-ribs	0.972 ± 0.007	0.974 ± 0.006	0.985 ± 0.004
	nnunet-kits	0.974 ± 0.016	0.977 ± 0.014	0.983 ± 0.010
	nnunet-thor	0.972 ± 0.017	0.974 ± 0.017	0.980 ± 0.012
	nnunet-liver	0.948 ± 0.040	0.953 ± 0.036	0.951 ± 0.043
	nnunet-abdmn	0.906 ± 0.062	0.916 ± 0.056	0.944 ± 0.040
	unetr	0.858 ± 0.076	0.888 ± 0.061	0.944 ± 0.033
	swin-unetr-small	0.497 ± 0.268	0.708 ± 0.185	0.932 ± 0.045
	nnunet-pancreas	0.769 ± 0.147	0.855 ± 0.082	0.926 ± 0.033
	swin-unetr-tiny	0.460 ± 0.274	0.680 ± 0.194	0.925 ± 0.048
	swin-unetr-base	0.010 ± 0.010	0.026 ± 0.026	0.819 ± 0.123
Hausdorff95	totsegm-vertebrae	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	totsegm-organs	1.41 ± 0.41	1.41 ± 0.41	1.00 ± 0.73
	totsegm-muscles	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	totsegm-all	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	nnunet-spleen	2.00 ± 1.00	1.73 ± 0.73	1.00 ± 0.00
	totsegm-cardiac	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	totsegm-ribs	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	nnunet-kits	1.73 ± 0.73	1.41 ± 0.41	1.41 ± 0.41
	nnunet-thor	1.41 ± 0.41	1.41 ± 0.41	1.00 ± 0.00
	nnunet-liver	11.64 ± 9.52	9.43 ± 8.02	11.87 ± 10.46
	nnunet-abdmn	3.61 ± 2.19	3.00 ± 1.59	2.24 ± 1.24
	unetr	4.12 ± 1.62	3.61 ± 1.37	2.00 ± 0.24
	swin-unetr-small	47.03 ± 37.18	19.12 ± 15.96	2.24 ± 0.82
	nnunet-pancreas	13.04 ± 9.04	9.35 ± 6.52	4.58 ± 2.35
	swin-unetr-tiny	59.01 ± 39.03	31.24 ± 24.24	3.00 ± 1.00
	swin-unetr-base	163.43 ± 64.73	147.99 ± 64.90	9.54 ± 7.30

5.2 Robustness across labels

Identifying challenging labels that contribute to the values presented in Table 2 would aid in addressing the robustness problem more specifically. However, the selected set of segmentation models lacks a common subset of labels applicable for straightforward analysis. Therefore, we employ label sizes to emphasize an overall trend. In Figure 5, we present Dice scores calculated between the 20% and 100% dose levels scans reconstructed using iterative denoising by the effective radius. The radius is derived based on the volume V occupied by the predicted label

$$r_{eff} = \sqrt[3]{\frac{3}{4\pi}V}. \tag{7}$$

The results show increased volatility and decreased value for smaller labels, including partially imaged organs. However, starting from the radii between 12–19 mm, robustness achieves higher and more stable values, although outliers are possible even for the largest organs. Despite labels not being characterized solely by their size, we consider the results as an early indication that models segmenting smaller lesions, which are beyond the scope of this study, may be less robust to dose reduction.

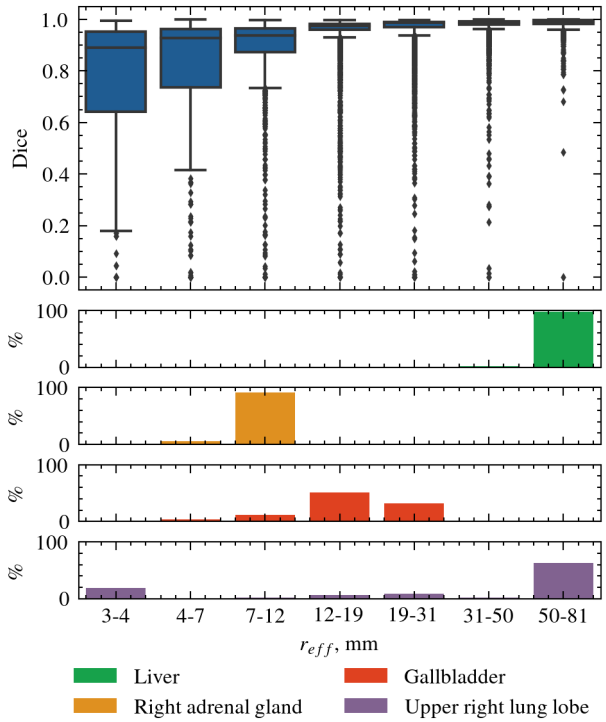


Figure 5: Distribution of Dice scores across effective radius intervals at the top along with the following binned effective radius histograms for different organs: the liver, the right adrenal gland, the gallbladder, the lung upper right lobe. The size of the intervals is sampled logarithmically. Accidental outliers of label size, e.g., for the lung upper right lobe, are caused by FOV clipping.

5.3 Contrast factor

Contrasted CT acquisitions help to better resolve structures inside the human body. Therefore, organ segmentations without contrast may be less robust to dose reduction since it affects the CNR. Figure 6 shows the distributions of Dice calculated between the 20% and 100% dose levels for reconstructions using iterative denoising for contrasted and non-contrasted CT scans. All the considered models show different levels of improvement for contrasted scans. However, for some models, non-contrasted scans lie outside the training data distribution. For example, *nnunet-kits* has the largest difference in robustness between contrasted and non-contrasted cases, but the training dataset mostly contains cases in the late arterial contrast phase. At the same time, the contrast factor has a minor impact on *totsegm-all* which was trained including non-contrasted CT scans. *totsegm-organs* is influenced the most by the contrast factor among five included in *totsegm-all*. Robustness of *totsegm-vertebrae*, *totsegm-muscles*, *totsegm-ribs* is not affected by contrast injection as they do not take up contrast.

5.4 Outliers

To account for metrics that lie at the extremes of the value range, the results were aggregated using the median and the median absolute deviation, as shown in Table 2. The outliers can also be observed in Fig. 5 and Fig. 6. The root cause of these outliers is twofold: false positives for small structures (such as the gallbladder, the adrenal gland) and FOV clipping. Both of these are a consequence of using a separate unlabeled dataset for the experiments, which may contain minor biases compared with the original training dataset. Since there is no ground truth available, the outliers cannot be adequately filtered. Therefore, we use more robust statistics to present the results.

5.5 Failure dynamics

Finally, we present failure curves (Fig. 7) for the least robust *swin-unetr-base* model and the *nnunet-abdmn* model, both trained using the same dataset. These curves show the Dice coefficient calculated for the considered dose levels. Since the training data are the same, both approaches share the same set of labels. In Figure 7a, for *nnunet-abdmn*, the decline in robustness is small and almost linear. The curves corresponding to the three considered reconstruction options overlap, although CNN-based denoising leads to more robust predictions.

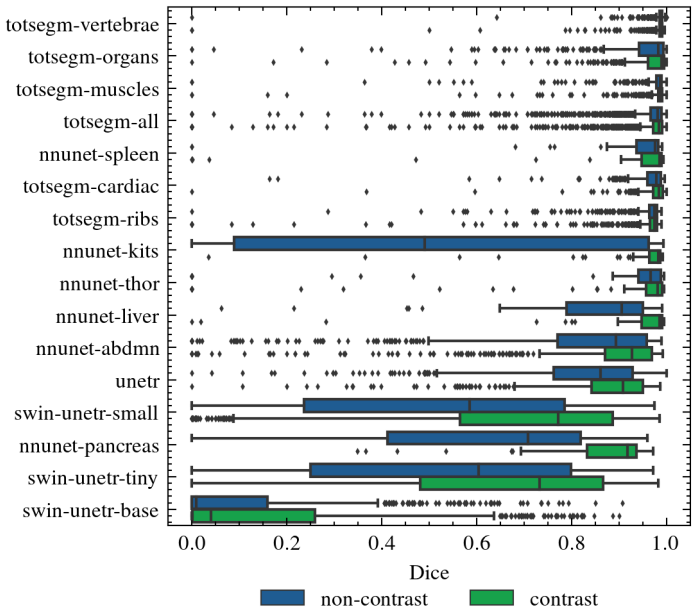


Figure 6: Distribution of Dice scores calculated between the 20% and 100% dose levels for reconstructions using iterative denoising.

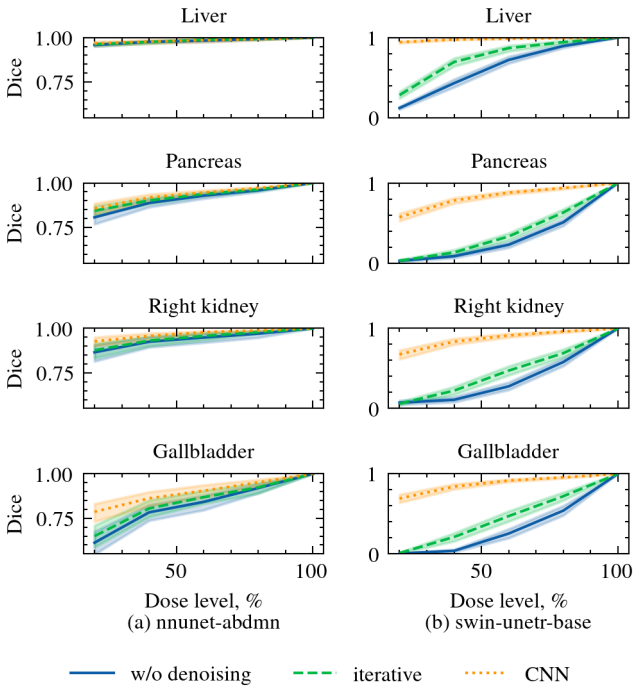


Figure 7: Dice scores calculated for different dose levels. The first column (a) corresponds to *nnunet-abdmn*, the second (b) - to *swin-unetr-base*. Four rows correspond to the four organs: the liver, the pancreas, the right kidney and the gallbladder.

The role of denoising becomes more significant for *swin-unetr-base* (Fig. 7b), where CNN-based denoising even changes the convexity of the curves compared with the other options. The described dynamics are consistent across all labels.

5.6 Limitations

This study has potential limitations. Our method assesses deviations from full-dose segmentation, not from the ground truth. Consequently, the accuracy of a segmentation method is decoupled from its robustness, as defined in this paper. For example, a constant function would exhibit complete robustness. Therefore, it is crucial to consider segmentation accuracy in conjunction with robustness. In this work, we employ pre-trained models, and their accuracy is evaluated and substantiated in the respective papers. Nevertheless, transitioning to the data used in our study may introduce a minor domain shift. Although visual results in supplementary Figures S1-S11 indicate adequate generalization, this may manifest as outliers observed in the results. This limitation can be mitigated by incorporating proper annotations for the data used to simulate low-dose images. Alternatively, another model for low-dose simulation can be explored. For example, a generative adversarial network (GAN) can be used to generate low-dose images from full-dose images. However, this noise model would require training data and additional validation to ensure that the generated images are realistic.

6 Conclusion

We introduced a method to test robustness of semantic segmentation models against dose reduction in CT images. Our approach allows systematic and practically relevant analysis with respect to various dose levels and intrinsically co-registered low- and full-dose image pairs without rescanning a patient. We analyzed pre-trained organ segmentation models in conjunction with different denoising options. Based on the results of the experiments, we can draw the following conclusions:

- Deep learning-based segmentation models for anatomical segmentation can exhibit robustness to dose reduction even without the application of additional denoising.
- The considered denoising methods do not compromise dose robustness of the segmentation models. Instead, they enable consistent segmentation results across a broader range of doses and can significantly change failure dynamics when reducing dose level.
- The use of a contrast agent usually improves robustness. This effect is more pronounced for the organs subjected to contrast.

Further research should address how to best couple image denoising and segmentation instead of developing them in a completely independent manner and which factors contribute to dose robustness the most. The proposed approach may also be extended with models segmenting lesions, which we expect to be less robust.

Disclosures

The authors have no relevant financial interests in the manuscript and no other potential conflicts of interest to disclose.

Data and Code Availability

The datasets generated and analyzed during the study are not publicly available due to privacy constraints. However, the derived data used for the analysis are available upon reasonable request. The code for the considered segmentation models is publicly available, and the corresponding repositories are mentioned in the text.

Acknowledgment

The authors are grateful to Stanislav Žabić giving access to the low-dose simulation software framework from the original work [42].

References

[1] Marco Aiello, Dario Baldi, Giuseppina Esposito, Marika Valentino, Marco Randon, Marco Salvatore, and Carlo Cavaliere. Evaluation of AI-Based Segmentation Tools for COVID-19 Lung Lesions on Conventional and Ultra-low Dose CT Scans. *Dose-Response*, 20(2):155932582210828, April 2022.

[2] I Arapakis, E Efstathopoulos, V Tsitsia, S Kordolaimi, N Economopoulos, S Argentos, A Ploussi, and E Alexopoulou. Using "iDose4" iterative reconstruction algorithm in adults' chest-abdomen-pelvis CT examinations: effect on image quality in relation to patient radiation exposure. *The British journal of radiology*, 87(1036):20130613, April 2014.

[3] Thorsten Buzug. *Computed Tomography*. Springer Berlin, Heidelberg, 2008.

[4] Hu Chen, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans. Medical Imaging*, 36(12):2524–2535, 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018.

[6] Francesco Ciompi, Kaman Chung, Sarah J. van Riel, Arnaud Arindra Adiyoso Setio, Paul K. Gerke, Colin Jacobs, Ernst Th. Scholten, Cornelia Schaefer-Prokop, Mathilde M. W. Wille, Alfonso Marchianò, Ugo Pastorino, Mathias Prokop, and Bram van Ginneken. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific Reports*, 7(1):46479, Apr 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE.

[8] Manoj Diwakar and Manoj Kumar. A review on CT image noise and its denoising. *Biomed. Signal Process. Control.*, 42:73–88, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[10] Mohammad Eslami, Solale Tabarestani, and Malek Adjouadi. Joint Low Dose CT Denoising And Kidney Segmentation. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*, pages 1–4, Iowa City, IA, USA, April 2020. IEEE.

[11] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015. https://pubs.rsna.org/doi/pdf/10.1148/radiol.2015151169.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Joël Greffier, Salim Si-Mohamed, Julien Frandon, Maeliss Loisy, Fabien de Oliveira, Jean Paul Beregi, and Djamel Dabli. Impact of an artificial intelligence deep-learning reconstruction algorithm for CT on image quality and potential dose reduction: A phantom study. *Medical physics*, 49(8):5052–5063, August 2022.

[14] Emily Hammond, Chelsea Sloan, John D. Newell, Jered P. Sieren, Melissa Saylor, Craig Vidal, Shayna Hogue, Frank De Stefano, Alexa Sieren, Eric A. Hoffman, and Jessica C. Sieren. Comparison of low- and ultralow-dose computed tomography protocols for quantitative lung and airway assessment. *Medical Physics*, 44(9):4747–4757, September 2017.

[15] 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.

[16] 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.

[17] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjan Sathianathen, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021.

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[19] Sarah M. Hooper, Jared A. Dunnmon, Matthew P. Lungren, Sanjiv Sam Gambhir, Christopher Ré, Adam S. Wang, and Bhavik N. Patel. Assessing robustness to noise: Low-cost head CT triage. *CoRR*, abs/2003.07977, 2020.

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[21] Zhenxing Huang, Zhou Liu, Pin He, Ya Ren, Shuluan Li, Yuanyuan Lei, Dehong Luo, Dong Liang, Dan Shao, Zhanli Hu, and Na Zhang. Segmentation-guided Denoising Network for Low-dose CT Imaging. *Computer Methods and Programs in Biomedicine*, 227:107199, December 2022.

[22] Icrp. Recommendations of the icrp. icrp publication 26. *Ann. ICRP*, 1(3):1–53, 1977.

[23] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021.

[24] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8825–8835. Computer Vision Foundation / IEEE, 2020.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[26] Florent Lalys, Simon Esneault, Miguel Castro, Lucas Royer, Pascal Haigron, Vincent Auffret, and Jacques Tomasi. Automatic aortic root segmentation and anatomical landmarks detection for tavi procedure planning. *Minimally invasive therapy & allied technologies*, 28(3):157–164, 2019.

[27] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Ruan. Segthor: Segmentation of thoracic organs at risk in CT images. *CoRR*, abs/1912.05950, 2019.

[28] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[29] Zhoubo Li, Lifeng Yu, Joshua D. Trzasko, David S. Lake, Daniel J. Blezek, Joel G. Fletcher, Cynthia H. McCollough, and Armando Manduca. Adaptive nonlocal means filtering based on local noise level for CT denoising: Adaptive nonlocal means filtering for CT denoising. *Medical Physics*, 41(1):011908, December 2013.

[30] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[31] Siqi Liu, Arnaud Arindra Adiyoso Setio, Florin C. Ghesu, Eli Gibson, Sasa Grbic, Bogdan Georgescu, and Dorin Comaniciu. No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting With Adversarial Attacks. *IEEE Transactions on Medical Imaging*, 40(1):335–345, January 2021.

[32] Armando Manduca, Lifeng Yu, Joshua D. Trzasko, Natalia Khaylova, James M. Kofler, Cynthia M. McCollough, and Joel G. Fletcher. Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT: Projection space denoising with bilateral filtering in CT. *Medical Physics*, 36(11):4911–4919, October 2009.

[33] Cynthia H. McCollough, Adam C. Bartley, Rickey E. Carter, Baiyu Chen, Tammy A. Drees, Phillip Edwards, David R. Holmes III, Alice E. Huang, Farhana Khan, Shuai Leng, Kyle L. McMillan, Gregory J. Michalak, Kristina M. Nunez, Lifeng Yu, and Joel G. Fletcher. Low-dose ct for the detection and classification of metastatic liver lesions: Results of the 2016 low dose ct grand challenge. *Medical Physics*, 44(10):e339–e352, 2017.

[34] Daniela Muenzel, Thomas Koehler, Kevin Brown, Stanislav Žabić, Alexander A. Fingerle, Simone Waldt, Edgar Bendik, Tina Zahel, Armin Schneider, Martin Dobritz, Ernst J. Rummeny, and Peter B. Noël. Validation of a low dose simulation technique for computed tomography images. *PLOS ONE*, 9(9):1–8, 09 2014.

[35] Konrad P Nesteruk, Mislav Bobić, Gregory C Sharp, Arthur Lalonde, Brian A Winey, Lena Nenoff, Antony J Lomax, and Harald Paganetti. Low-dose computed tomography scanning protocols for online adaptive proton therapy of head-and-neck cancers. *Cancers (Basel)*, 14(20):5155, October 2022.

[36] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick Ferdinand Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William R. Jarnagin, Maureen McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR*, abs/1902.09063, 2019.

[37] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20698–20708, New Orleans, LA, USA, June 2022. IEEE.

[38] A. Tsanda, H. Nickisch, T. Wissel, T. Klinder, T. Knopp, and M. Grass. On TotalSegmentator's performance on low-dose CT images. In Olivier Colliot and Jhimli Mitra, editors, *Medical Imaging 2024: Image Processing*, volume 12926, page 129260B. International Society for Optics and Photonics, SPIE, 2024.

[39] Jing Wang, Tianfang Li, Hongbing Lu, and Zhengrong Liang. Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography. *IEEE Trans. Medical Imaging*, 25(10):1272–1283, 2006.

[40] Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[41] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K. Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Trans. Medical Imaging*, 37(6):1348–1357, 2018.

[42] Stanislav Žabić, Qiu Wang, Thomas Morton, and Kevin M. Brown. A low dose simulation tool for CT systems with energy integrating detectors: A low dose simulation tool. *Medical Physics*, 40(3):031102, February 2013.

Artyom Tsanda is a PhD student at Hamburg University of Technology, working at the Institute for Biomedical Imaging headed by Prof. Dr.-Ing. Tobias Knopp. His research interests include the development of deep learning algorithms for medical image reconstruction.

Biographies and photographs of the other authors are not available.

List of Figures

1 A full-dose CT scan of the phantom used to validate the low-dose simulation along with the selected ROIs. The sampled standard deviation $\hat{\sigma}$ is calculated for each ROI and for each dose level. The deviation across dose levels for each ROI is shown next to the image.

2 Examples of reconstructed images at 20% dose level with different denoising methods applied: (a) without, (b) iterative and (c) CNN. Level/Window is set to 50/500.

3 A segmentation example for *totalsegm-all* calculated for (b) 20% and (a) 100% dose level images with iterative denoising applied. The last image (c) shows the difference between the two segmentations.

4 Kernel density estimations of sampled standard deviations calculated inside the liver region for each CT scan from different datasets: 20% and 100% dose data reconstructed using iterative denoising and TotalSegmentator training data.
5 Distribution of Dice scores across effective radius intervals at the top along with the following binned effective radius histograms for different organs: the liver, the right adrenal gland, the gallbladder, the lung upper right lobe. The size of the intervals is sampled logarithmically. Accidental outliers of label size, e.g., for the lung upper right lobe, are caused by FOV clipping.
6 Distribution of Dice scores calculated between the 20% and 100% dose levels for reconstructions using iterative denoising.
7 Dice scores calculated for different dose levels. The first column (a) corresponds to *nnunet-abdmn*, the second (b) - to *swin-unetr-base*. Four rows correspond to the four organs: the liver, the pancreas, the right kidney and the gallbladder.

List of Tables

1 Taxonomy of segmentation models included into the study.
2 Differences between segmentation masks calculated for the corresponding 20% and 100% dose level CT scans. The difference is measured with the Dice score and the Hausdorff distance (95 percentile). The median and the median absolute deviation is calculated across all CT images for a given pair of a model and denoising method.