

Latent-to-Latent Flow for Volumetric Stochastic Segmentation

Omar Todd¹, Sooha Kim², Raghav Mehta¹, Katherine Mackay³, David Bernstein², Alexandra Taylor², Fabio De Sousa Ribeiro^{1,†}, and Ben Glocker^{1,†}

¹ Imperial College London, UK

² The Royal Marsden Hospital, London, UK

³ Imperial College Healthcare NHS Trust, London, UK
omar.todd16@imperial.ac.uk

Abstract. Uncertainty arising from inter-observer variability in medical image segmentation plays an important role in developing treatment plans. Research in this area is inhibited by the lack of multiple annotations for large-scale medical datasets, especially for volumetric data, which suffers from additional scaling and computational complexity challenges. Flow matching has emerged as a powerful framework for generative modelling and has also been demonstrated to maintain strong performance when working with latent representations of images. In this work, we introduce a latent-to-latent flow technique for stochastic segmentation of medical volumes via encoded representations of both the image and label space. We evaluate our method on two challenging applications covering delineation uncertainty for radiotherapy planning and multiple organ structure segmentation, improving efficiency up to 14× compared with full resolution models while maintaining clinically relevant performance. Code available at: <https://github.com/biomed-mira/L2L-Flow>.

Keywords: Volumetric Segmentation · Uncertainty · Flow Matching

1 Introduction

Medical image segmentation aims to provide accurate contouring of anatomical structures to aid clinicians in developing treatment plans. Segmenting medical images is an inherently ambiguous task due to uncertainty in the exact positions of organ and tissue boundaries. Incorporating uncertainty into the planning process is important for improving patient outcomes [19,12]. Uncertainty in machine learning is broadly categorized into two classes: epistemic, which concerns the model, and aleatoric, which concerns the data [28]. Inter-observer variation (IOV) [20], arising from clinical experts contouring a structure differently, is an important source of aleatoric uncertainty that deterministic segmentation models, such as U-Net [23], are unable to accurately capture. This requires the use of stochastic segmentation models, which learn a probability distribution over the plausible segmentation maps to simulate clinically meaningful uncertainty.

[†] Joint senior authors.

While large amounts of medical imaging data are represented as 3D volumes, much of the stochastic segmentation research and benchmarking is developed for 2D datasets [2,6]. Data and label scarcity is a prevalent issue due to the labour-intensive nature of collecting multiple expert annotations, which is exacerbated for 3D volumes. Additionally, probabilistic modelling of volumetric annotations comes with other challenges such as increased computational complexity. This can be an issue for time-sensitive applications such as adaptive radiotherapy planning. Recently, flow matching has emerged as an attractive framework for generative modelling on natural images [15,16,1]; however, its suitability for volumetric stochastic segmentation has not been explored and is the main focus of this work. Our primary contributions are as follows:

- (i) We identify performance degradation of Flow-SSNs [7] on high-resolution volumetric medical data, and introduce a bespoke time-shifted noise schedule resulting in significantly improved stability and performance, setting a new benchmark for volumetric stochastic segmentation.
- (ii) We introduce the Latent-to-Latent Flow (L2L-Flow), an efficient and scalable model alternative to our time-shifted Flow-SSN, allowing for a wider range of volumetric clinical applications. Notably, L2L-Flow provides a controllable efficiency/performance trade-off, demonstrating up to $14\times$ faster inference compared to Flow-SSNs whilst retaining highly competitive performance.
- (iii) We evaluate our flow models on two challenging medical datasets for radiotherapy planning and multi-organ segmentation, demonstrating clinically relevant uncertainty estimation capabilities, and highlighting the potential of generative approaches for volumetric segmentation under label ambiguity.

2 Related Work

Probabilistic U-Net [13] utilized a U-Net alongside a conditional variational auto-encoder to learn a conditional density function for generating stochastic segmentations. PhiSeg [3] built on this model by introducing a hierarchical probabilistic model for segmenting the image via latent distributions at different resolutions. Stochastic Segmentation Networks (SSNs) [21] explored limitations of the pixel-wise independence assumed in previous work, and employed a more expressive low-rank Gaussian distribution for modelling pixel variance. This model, however, is prone to stability issues during training, and the low-rank approximation imposes limitations on the complexity of the distribution it can learn [7]. More recently, Mixture-of-Stochastic-Experts (MoSE) modelled each annotator as a separate distribution, from which samples were optimized via an optimal transport loss function [9]. For 3D data, this model can suffer from scaling issues if it is desired to have a high number of samples per mode during training.

Diffusion models are another generative model class that have been successfully applied to the task of stochastic segmentation in 2D [26,27]. While they have demonstrated the ability to generate high-quality segmentations, Denoising Diffusion Probabilistic Model (DDPM) [10] based methods are very expensive at inference time, which becomes increasingly undesirable as the data scale

increases. More efficient options such as Denoising Diffusion Implicit Models (DDIM) [24] can help with this; however, the paths induced by these methods often have a complex shape that is difficult to solve. Most recently, the flow matching paradigm has emerged as a simple yet powerful setup for generative modelling [15,16,1]. Flow-SSNs [7] introduced a modified rectified flow with a learnable image-conditioned prior in one-hot label space and a categorical likelihood for medical stochastic segmentation. This approach showed strong performance in 2D, while providing vastly superior efficiency compared to diffusion-based stochastic segmentation models [27]. Recently, latent rectified flows [8] have also gained traction. Ma et al. [17] applied them to cardiac segmentations to facilitate data augmentation for shape completion. Focusing on stochastic segmentation, however, it is crucial for the latent flow to utilize the rich structural information in the high-resolution images and labels to preserve strong clinical utility when modelling the distribution in a lower-dimensional space.

3 Methodology

3.1 Neural ODEs and Flow Matching

Continuous Normalising Flows (CNFs) [5] learn a mapping from a source $\mathbf{x}_0 \sim p_{\text{src}}$ to a target distribution $\mathbf{x}_1 \sim p_{\text{tgt}}$ via an ordinary differential equation (ODE)

$$d\mathbf{x}_t = v_t(\mathbf{x}_t; \theta) dt, \quad t \in [0, 1], \quad (1)$$

where v_t is a time-dependent velocity field parameterised by a neural network.

Flow Matching [15,16,1] trains CNFs without simulating trajectories by specifying a conditional probability path between $\mathbf{x}_0$ and $\mathbf{x}_1$, such as simple linear interpolation $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, yielding a least-squares regression objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \left[\|v_t(\mathbf{x}_t; \theta) - v_t^*(\mathbf{x}_t | \mathbf{x}_1)\|^2 \right], \quad (2)$$

where $t \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0 \sim p_{\text{src}}$ and $\mathbf{x}_1 \sim p_{\text{tgt}}$. Flow-SSNs [7] learn a mapping from an image conditional prior $\mathbf{u}|\mathbf{x} \sim p_{\text{src}}$ to a label map distribution $\mathbf{y} \sim p_{\text{tgt}}$. Since targets are discrete, Flow-SSNs use a linear interpolant in one-hot space $\mathbf{y}_t = (1-t)\mathbf{u} + t\mathbf{y}$, and an expected categorical likelihood: $\max_{\lambda, \theta} \mathbb{E}_{t, \mathbf{u}, \mathbf{y}} [\log p(\mathbf{y} | \mathbf{y}_t; \theta)]$, where the source $\mathbf{u} \sim p_{\text{src}}(\mathbf{u}|\mathbf{x}; \lambda)$ is learnt from data rather than being fixed.

3.2 Time-Shifted Noise Scheduling

Recent work on modern flows showed that the use of a shifted noise schedule leads to improved sample quality for high-resolution image data [14]. We find that careful calibration of the noise schedule is critical to the effective use of Flow-SSNs in high-dimensional volumetric segmentation tasks, which has been previously overlooked [7]. We propose a modified version of the shift schedule [14], which accounts for our path direction going from $t = 0$ to $t = 1$, yielding:

$$\mathbf{y}_t = (1 - t_s)\mathbf{u} + t_s\mathbf{y}, \quad \text{where} \quad t_s = \frac{t}{\alpha - t(\alpha - 1)}. \quad (3)$$

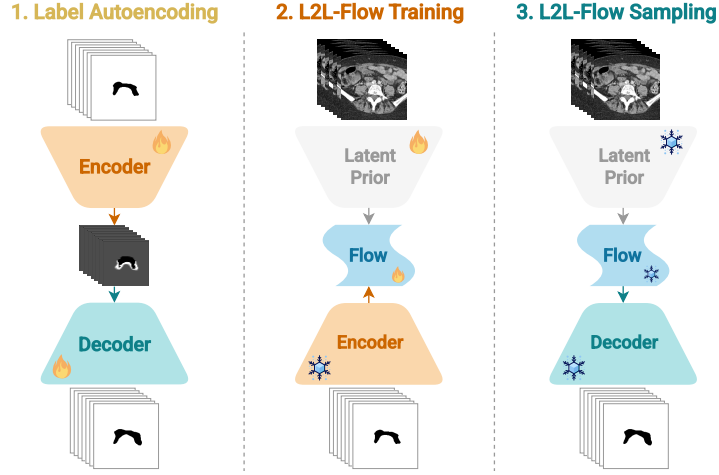


Fig. 1. Architectural overview of L2L-Flow. (*Left*) A volumetric segmentation label autoencoder is trained to obtain latent label representations. (*Middle*) A latent flow is then trained to learn a transport from a conditional prior given input volumes to frozen latent label maps. (*Right*) Stochastic segmentation is performed by integrating the ODE, mapping conditional samples from the volume prior to latent label predictions. Finally, the ODE solution is upsampled to full resolution using the label decoder.

where higher values of the hyperparameter α induce more aggressive noising. As will be demonstrated, this simple modification unlocks significant training stability and performance improvements for Flow-SSNs in volumetric settings.

3.3 Latent-to-Latent Flow

Unlike standard latent flows/diffusions which model the distribution of a single high-dimensional image variable $p(\mathbf{x})$, generative segmentation seeks to estimate the label conditional distribution $p(\mathbf{y} | \mathbf{x}) = \int p(\mathbf{y}, \mathbf{z} | \mathbf{x}) d\mathbf{z}$, where both $\mathbf{x}$ and $\mathbf{y}$ are high-dimensional, and $\mathbf{z}$ is a latent variable. Flow-SSNs [7] model this joint in observation space, making sampling too computationally expensive for volumes. To overcome this, we propose Latent-to-Latent Flow (L2L-Flow), a new scalable approach that learns a flow transport from a *latent* volume-conditional prior to a *latent* label distribution. As described in Fig. 1, the first stage consists of training a volumetric segmentation label autoencoder $D_\psi(E_\phi(\mathbf{y}))$ to obtain compressed label latents $\mathbf{z}_1 = E_\phi(\mathbf{y})$. We then define the latent linear flow interpolant:

$$\mathbf{z}_t = (1 - t)E_\lambda(\mathbf{x}) + tE_\phi(\mathbf{y}), \quad t \in [0, 1], \quad (4)$$

where *both* the source and target distributions are parameterised by encoders:

$$p_{\text{src}}(\mathbf{z}_0 | \mathbf{x}; \lambda) = \mathcal{N}(E_\lambda(\mathbf{x}), I), \quad \text{and} \quad p_{\text{tgt}}(\mathbf{z}_1 | \mathbf{y}; \phi) = \delta(\mathbf{z}_1 - E_\phi(\mathbf{y})). \quad (5)$$

Dissimilar to Flow-SSNs, the latent target variable $\mathbf{z}_1 = E_\phi(\mathbf{y})$ here is no longer a discrete one-hot label map, so a least-squares regression objective can be used:

$$\min_{\lambda, \theta} \mathbb{E}_{t, \mathbf{z}_0 \sim p_{\text{src}}, \mathbf{z}_1 \sim p_{\text{tgt}}} \left[\|v_t(\mathbf{z}_t; \theta) - (E_\phi(\mathbf{y}) - \mathbf{z}_0)\|^2 \right], \quad (6)$$

noting that $v_t(\mathbf{z}_t; \theta)$ and $E_\lambda(\mathbf{x})$ are trained simultaneously, while we find that pretraining and freezing $E_\phi(\mathbf{y})$ leads to improved stability and performance.

At inference time, stochastic segmentation is performed by: (i) sampling from the latent volume-conditional prior $\mathbf{z}_0 \sim p_{\text{src}}(\mathbf{z}_0 \mid \mathbf{x}; \lambda)$; (ii) solving the flow ODE by integrating $\mathbf{z}_1 = \mathbf{z}_0 + \int_0^1 v_t(\mathbf{z}_t; \theta) dt$; and (iii) upsampling the ODE solution using the frozen pretrained label decoder $\mathbf{y} = D_\psi(\mathbf{z}_1)$. As we will demonstrate empirically, L2L-Flow accelerates flow sampling by up to $14\times$ while maintaining volumetric segmentation performance comparable to full-resolution models.

4 Experiments

4.1 Datasets

To address the challenges of gathering high-quality IOV for medical data, recent works have researched more efficient annotation schemes using clinician-defined ranges (CDR) [4, 18, 25]. These ranges provide soft inner and outer boundaries within which the true contour is expected to be positioned, while indicating the non-uniform, spatially varying uncertainty of the contour shape. To explore the suitability of CDRs for stochastic segmentation, we evaluate the clinical relevance of samples from probabilistic models trained on our private CT radiotherapy dataset of clinical target volumes (CTV). We also evaluate our method on the publicly available multi-class CURVAS challenge dataset with three annotations from multiple raters for three structures: pancreas, kidney, and liver [22].

4.2 Implementation Details

The radiotherapy and CURVAS datasets used for evaluation in this work had a limited number of cases (55 and 90 cases, respectively). As such, we utilize a 5-fold cross-validation strategy for evaluation. We use the standard generalised energy distance (GED) and diversity⁴ with 16 samples to measure the performance of stochastic segmentation methods [13, 21, 7, 9].

For both datasets, the high-resolution volumes of interest were taken as 192^3 ROI crops of the images. Each model was trained using a batch size of 1 with gradient accumulation of 4 steps to accommodate the computational costs of modelling the distribution on large volumes. The models were all trained using a learning rate of $1e-4$ for 2000 epochs. All experiments were performed using an NVIDIA L40 (48GB) GPU. The flow matching components of the models

⁴ Recall that diversity is only relevant contextually, as high diversity can be trivially achieved with random noise as a model.

Table 1. Quantitative cross-validation results on the Radiotherapy dataset. L2L-Flow demonstrates highly competitive segmentation performance compared to full resolution models, while being more than 14 $\times$ faster than Flow-SSN during inference.

METHOD	$D_{\text{GED}}^2 \downarrow$	Diversity $\uparrow$	Dice $\uparrow$	s/Img $\downarrow$	Params
MoSE [9]	0.183 $\pm$.008	0.137 $\pm$.005	0.893 $\pm$.004	0.824 $\pm$.010	6.9M
PhiSeg [3]	0.177 $\pm$.007	0.122 $\pm$.002	0.898 $\pm$.003	4.895 $\pm$.013	63M
SSN [21]	0.167 $\pm$.005	0.160 $\pm$.007	0.896 $\pm$.003	0.622 $\pm$.010	7.1M
Flow-SSN [7]	0.165 $\pm$.008	0.168 $\pm$.008	0.894 $\pm$.004	15.296 $\pm$.013	6.9M
Flow-SSN (shifted)	0.155 $\pm$.005	0.164 $\pm$.008	0.899 $\pm$.001	13.471 $\pm$.011	6.9M
L2L-Flow	0.161 $\pm$.004	0.182 $\pm$.008	0.896 $\pm$.002	0.936 $\pm$.016	9.3M

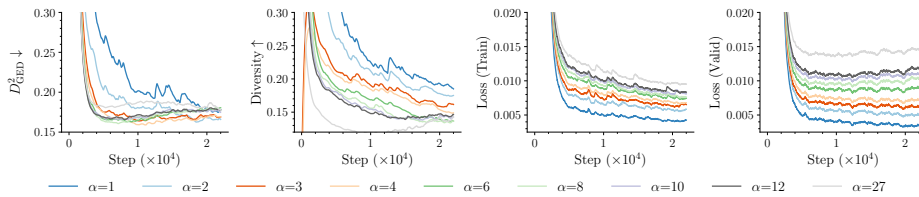


Fig. 2. Ablation of shifted noise schedules. In high-dimensional volumetric settings, we observe significant performance improvements to GED by shifting the Flow-SSN noise schedule. Higher values of α are more prone to overfitting during training.

described in this work use Residual U-Net backbones [7] and use 8 time steps for ODE solving during inference. For the Flow-SSN using a shifted noise schedule, we found $\alpha=3$ (see Fig. 2) to give the best results in our experiments. For L2L-Flow, on the radiotherapy dataset, we down-sample images by a factor of 4 for a latent resolution of 48^3 . On the CURVAS dataset, we additionally down-sample images by a factor of 2 to a 96^3 resolution. This is achieved via varying the number of levels in the U-Net and the auto-encoder.

4.3 Radiotherapy Delineation Uncertainty

Tab. 1 shows the average performance across folds of L2L-Flow alongside each of the baseline methods. For the full-resolution stochastic models, our time-shifted Flow-SSN provides the strongest overall results. We observe that L2L-Flow demonstrates highly competitive performance for GED compared with the full-resolution flow model while being $\approx 14\times$ more efficient. A qualitative evaluation to better discern the clinical relevance of the best-performing models is shown in Fig. 3. This was performed for a case where ground-truth IOV was available from 10 clinical experts. The expert contours showed numerous overlaps and crossings, which are reflected in the variability of samples from the flow-based models. However, despite strong quantitative performance, the SSN contours only varied in a strong concentric manner. From a clinical standpoint, the strong dissimilarity of these contours to the randomness of human IOV made

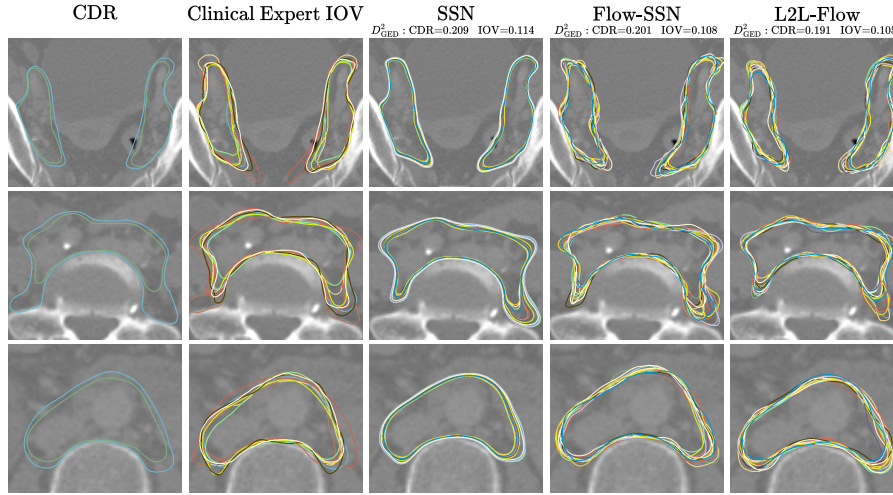


Fig. 3. Qualitative results on the Radiotherapy dataset. Stochastic segmentation samples are compared with a representative example of human expert IOV. GED is reported for samples relative to both the CDR and IOV. Crucially, unlike SSN, L2L-Flow and Flow-SSN show clinically relevant variation compared to expert contours.

them unsuitable for downstream planning. This distinction is not captured solely by the GED metric and highlights the need for more holistic comparisons when incorporating machine learning models into the clinical workflow.

4.4 Multi-Organ Segmentation (CURVAS Dataset)

In the original challenge, the best-performing model was the deterministic nnU-Net [11] which we include as an additional baseline. Per-structure quantitative results are shown in Fig. 4. Our time-shifted Flow-SSN performs the best, most notably with 3.4% better GED for the pancreas compared to the original Flow-SSN, albeit with substantially slower speeds compared to other models. The more sizeable performance gap on the pancreas between L2L-Flow with 4x downsampling and the best-performing Flow-SSN arises from data compression on an already small structure. To emphasize the flexibility of our model in terms of the efficiency/performance trade-off, we also report results for L2L-Flow at 2x downsampling. This increased performance results from better capturing fine details of the pancreas as shown in Fig. 5. Despite the increased inference time, this model still remains $\approx 5\times$ faster than the full resolution Flow-SSN. These experiments also highlighted the non-trivial difficulties of moving existing frameworks to 3D. These challenges can be both methodological, such as SSN’s low-rank sampling instability (cf. Fig. 5), and computational, such as with MoSE’s reduced sampling size during training due to memory constraints.

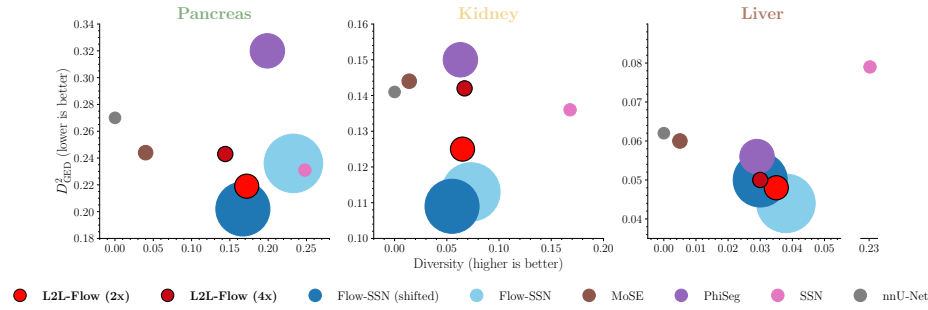


Fig. 4. Quantitative results on the CURVAS dataset. GED and diversity scores are reported, with circle size proportional to inference time. Flow-SSN (shifted) displays much better performance than the original for the most challenging pancreas class. Reducing the down-sampling used in L2L-Flow from $4\times$ to $2\times$ helps bridge the performance gap with full-resolution models, with a trade-off for slower inference time.

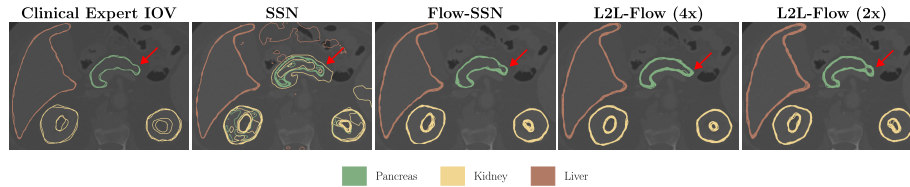


Fig. 5. Qualitative results on the CURVAS dataset. SSN produces spurious correlations for multi-class tasks. L2L-Flow at a higher latent resolution can capture more subtle structural properties compared with full-resolution counterparts. Red arrows indicate the $4\times$ down-sampled model failing to precisely segment a part of the pancreas.

5 Conclusion

We introduce L2L-Flow, a scalable generative approach for stochastic segmentation of volumetric data. The proposed model produces clinically meaningful samples while being substantially more efficient in both memory and computation than its full-resolution counterpart. This is achieved by leveraging high-level image and label information within a latent rectified flow formulation. Results across multiple datasets demonstrate that extending complex probabilistic models from 2D to 3D is non-trivial, evidenced by the methodological and training stability issues of prior methods. The radiotherapy experiments, in particular, expose limitations of current stochastic segmentation metrics when used in isolation and underscore the importance of complementing quantitative evaluation with qualitative and clinical considerations when developing machine learning models for clinical workflows. Future work will explore the integration of more expressive image and label encoders, such as those derived from foundation models, to further enhance the capabilities of L2L-Flow.

Acknowledgment

This work was supported by the European Union’s Horizon Europe research and innovation programme for the AI-POD project under grant agreement 101080302. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them. Additional support was received from the UKRI AI programme, and the EPSRC, for CHAI-EPSRC Causality in Healthcare AI Hub (grant no. EP/Y028856/1) and the Royal Academy of Engineering as part of the Kheiron/RAEng Research Chair. S.K. is supported by the CRUK Convergence Science Centre at The Institute of Cancer Research, London, and Imperial College London (A26234).

Disclosure of interests

B.G. is a part-time employee of DeepHealth. No other competing interests.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics* **38**(2), 915–931 (2011)
3. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötter, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 119–127. Springer (2019)
4. Bernstein, D., Taylor, A., Nill, S., Oelfke, U.: New target volume delineation and ptv strategies to further personalise radiotherapy. *Physics in Medicine & Biology* **66**(5), 055024 (2021)
5. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. *Advances in neural information processing systems* **31** (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. De Sousa Ribeiro, F., Todd, O., Jones, C., Kori, A., Mehta, R., Glocker, B.: Flow stochastic segmentation networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14754–14765 (2025)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)

9. Gao, Z., Chen, Y., Zhang, C., He, X.: Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. In: The Eleventh International Conference on Learning Representations (2023)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
12. Kinoshita, N., Shimizu, M., Motegi, K., Tsuruta, Y., Takakura, T., Oguchi, H., Kurokawa, C.: Quantification of uncertainties in reference and relative dose measurements, dose calculations, and patient setup in modern external beam radiotherapy. *Radiological physics and technology* **18**(1), 58–77 (2025)
13. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems* **31** (2018)
14. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
16. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
17. Ma, Q., Meng, Q., Qiao, M., Matthews, P.M., O'Regan, D.P., Bai, W.: Cardiacflow: 3d+ t four-chamber cardiac shape completion and generation via flow matching. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 89–99. Springer (2025)
18. Mackay, K., Bernstein, D., Kim, S., Glocker, B., Yasar, B., Cheng, L., Robinson, S.D., Khoo, V., Taylor, A.: 1621 assessment of a novel approach for validating auto-contouring systems with an expert predicted acceptability range. *Radiotherapy and Oncology* **206**, S2439–S2442 (2025)
19. McCrindle, B., Zukotynski, K., Doyle, T.E., Noseworthy, M.D.: A radiology-focused review of predictive uncertainty for ai interpretability in computer-assisted segmentation. *Radiology: Artificial Intelligence* **3**(6), e210031 (2021)
20. McErlean, A., Panicek, D.M., Zabor, E.C., Moskowitz, C.S., Bitar, R., Motzer, R.J., Hricak, H., Ginsberg, M.S.: Intra-and interobserver variability in ct measurements in oncology. *Radiology* **269**(2), 451–459 (2013)
21. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., Van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems* **33**, 12756–12767 (2020)
22. Riera-Marin, M., Rodriguez-Comas, J., May, M.S., Pan, Z., Zhou, X., Liang, X., Erick, F.X., Prenner, A., Hemon, C., Boussot, V., et al.: Calibration and uncertainty for multirater volume assessment in multiorgan segmentation (curvas) challenge results. *arXiv preprint arXiv:2505.08685* (2025)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)

24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
25. Todd, O., Kim, S., Mackay, K., Mehta, R., Ribeiro, F.D.S., Bernstein, D., Taylor, A., Glocker, B.: Delineation uncertainty from clinician ranges in cervical cancer radiotherapy planning. In: MICCAI Workshop on Deep Generative Models. pp. 82–92. Springer (2025)
26. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International conference on medical imaging with deep learning. pp. 1336–1348. PMLR (2022)
27. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1119–1129 (2023)
28. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology* **1**(1), 100003 (2023)