



FlowSDF: Flow Matching for Medical Image Segmentation Using Distance Transforms

Lea Bogensperger¹ · Dominik Narnhofer² · Alexander Falk¹ · Konrad Schindler² · Thomas Pock¹

Received: 28 May 2024 / Accepted: 4 February 2025 / Published online: 22 March 2025
© The Author(s) 2025

Abstract

Medical image segmentation plays an important role in accurately identifying and isolating regions of interest within medical images. Generative approaches are particularly effective in modeling the statistical properties of segmentation masks that are closely related to the respective structures. In this work we introduce FlowSDF, an image-guided conditional flow matching framework, designed to represent the signed distance function (SDF), and, in turn, to represent an implicit distribution of segmentation masks. The advantage of leveraging the SDF is a more natural distortion when compared to that of binary masks. Through the learning of a vector field associated with the probability path of conditional SDF distributions, our framework enables accurate sampling of segmentation masks and the computation of relevant statistical measures. This probabilistic approach also facilitates the generation of uncertainty maps represented by the variance, thereby supporting enhanced robustness in prediction and further analysis. We qualitatively and quantitatively illustrate competitive performance of the proposed method on a public nuclei and gland segmentation data set, highlighting its utility in medical image segmentation applications.

Keywords Flow matching · Image segmentation · Conditional generative models · Signed distance function

1 Introduction

Medical image segmentation is the process of assigning each pixel in an image to a semantic class, which is crucial for extracting valuable insights for diagnosis and treatment plan-

ing. Traditionally, this task was handled by methods that have now been largely superseded by deep neural networks. These networks are particularly effective due to their ability to learn discriminatively from large data sets in an end-to-end manner (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Wang et al., 2020). In addition to these discriminative models, generative models like generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have emerged for image segmentation with the advantage of learning the underlying statistics of segmentation masks conditioned on the respective images of interest (Iqbal et al., 2022; Xun et al., 2022; Amit et al., 2021). A very recently introduced generative approach is flow matching (Lipman et al., 2022; Liu et al., 2022), which uses an ordinary differential equation (ODE) to model the probability flow that transforms a data distribution into a simple, tractable prior distribution.

Furthermore, the process of modeling a probability flow can be extended to modeling the flow of a conditional probability, such as that of a segmentation mask given an image $p(m|x)$. When modeling the conditional distribution using diffusion/flow models, noise is injected on the segmentation masks to transform them into samples from a Gaussian prior distribution. The statistical properties of these masks

Communicated by Ullrich Köthe.

Lea Bogensperger and Dominik Narnhofer have contributed equally to this work.

✉ Lea Bogensperger
lea.bogensperger@tugraz.at

Dominik Narnhofer
dnarnhofer@ethz.ch

Alexander Falk
falk@tugraz.at

Konrad Schindler
schindler@ethz.ch

Thomas Pock
thomas.pock@tugraz.at

¹ Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria

² Photogrammetry and Remote Sensing, ETH Zürich, Zurich, Switzerland

are complex – typically bimodal or with only few modes based on the number of semantic classes involved. It can be challenging to learn these statistics, as there is no transition between class modes, which can result in unnatural distortions.

A promising approach to mitigate these issues utilizes implicit segmentation maps, represented by signed distance functions (SDF). An SDF at any point on the segmentation map computes the orthogonal distance to the nearest boundary, assigning a negative sign to points inside the boundary and a positive sign to those outside. This method not only facilitates a smoother distribution of segmentation masks but also ensures more natural transitions between class modes. Moreover, the use of SDFs promotes smoothness in the segmentation maps, which are derived by thresholding the SDF at object boundaries.

To this end, we introduce FlowSDF, an approach that combines flow matching with the SDF for generative medical image segmentation. We build on prior work of (Bogensperger et al., 2023) where a variance-exploding (VE) stochastic differential equation (SDE) is employed to also model a conditional diffusion process to learn the statistics of segmentation masks given by the SDF. We improve upon our previous work in (Bogensperger et al., 2023) by using a generative flow matching framework (Lipman et al., 2022) in place of the generative diffusion model. Our proposed image-guided flow matching method also learns a smooth implicit representation of segmentation masks conditioned on the corresponding images, see Fig. 1. This integration allows the noise-injection perturbation process to canonically smooth the distribution of SDF masks as further illustrated in Fig. 4. Additionally, FlowSDF leverages its generative nature to produce multiple segmentations for a given input image, thereby allowing for segmentation uncertainty quantification. This also enhances the robustness and interpretability of our approach and provides valuable insights into the segmentation process. Summarized, our main contributions are:

- To the best of our knowledge, we are the first to apply conditional flow matching for image segmentation.
- Further we combine this approach with distance transforms to obtain a smoother, implicit representation of segmentation masks.
- Throughout multiple experiments we achieve competitive results for the considered data sets.

2 Related Work

2.1 Medical Image Segmentation

In the following section, we discuss selected literature on medical image segmentation related to the proposed approach. Historically, techniques such as level set methods have been prominent (Osher and Sethian, 1988; Osher et al., 2004), which use a signed function to compute distance metrics to the nearest object, whose actual contour is represented by the zero level. Hence, the aim is to provide a precise spatial context of the object of interest by focusing on the respective boundary. Another notable method in the field of variational methods is the Potts model, also known as piecewise constant Mumford-Shah model (Chan and Vese, 2001; Mumford and Shah, 1989). It has been used extensively to handle segmentations by minimizing an energy function that discourages discontinuities between neighboring pixels unless there is a strong gradient in the image data, effectively enforcing smoothness while adapting to inherent image contours. Undoubtedly, there are many more classical techniques for image segmentation, such as active contours (Kass et al., 1988) and graph cuts (Boykov and Kolmogorov, 2004; Boykov and Funka-Lea, 2006), to name but a few.

The emergence of deep learning led to many discriminative methods that have significantly advanced the field of image segmentation using deep neural networks (Ronneberger et al., 2015; Badrinarayanan et al., 2017). These methods are designed to work end-to-end and have bene-

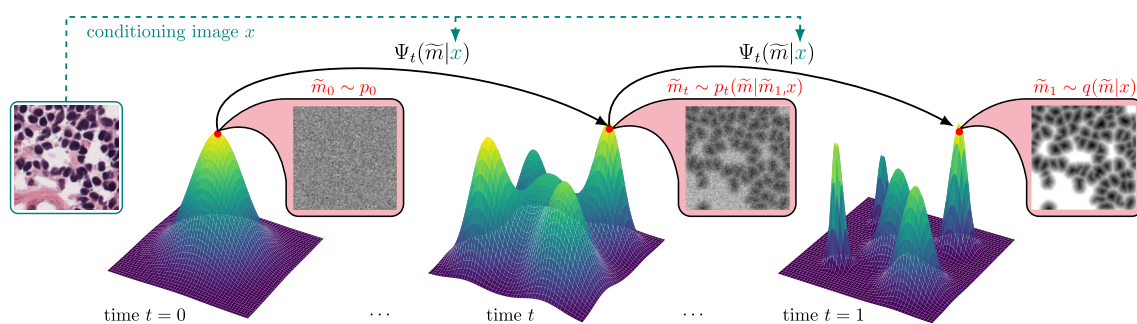


Fig. 1 Overview of FlowSDF, which models a conditional flow $\Psi_t(\tilde{m}|x)$ by relating a learned conditional vector field v_θ to the conditional probability path $p_t(\tilde{m}|\tilde{m}_1, x)$. During inference, a sample

$\tilde{m}_0 \sim p_0$ is drawn from a known prior distribution (left) and the ODE is numerically solved to obtain segmentation masks $\tilde{m}_1$ that resemble those from the conditional distribution $q(\tilde{m}|x)$ (right)

fited greatly from continual advancements in architecture search, allowing for more refined and effective models (Wang et al., 2020; Isensee et al., 2021; Cao et al., 2022). However, also some of these newer methods have recognized the vital information that can be extracted from information such as contours and boundaries, in particular for image segmentation. Therefore, there have been approaches incorporating shape priors and distance transforms for discriminative image segmentation using CNNs (Naylor et al., 2018; Xue et al., 2020; Brissman et al., 2021; Yu et al., 2023). Although these discriminative methods tend to be fast, easy to deploy, and yield good results, they often exhibit overconfidence in their predictions, leading to issues with calibration where the predicted probabilities do not align with actual outcomes.

Furthermore, when annotation data is available from multiple annotators to account for ambiguity in segmentation data, discriminative models can be transformed into probabilistic variants by incorporating stochastic information by sampling from a learned latent distribution using a conditional variational autoencoder (VAE) Kohl et al. (2018); Bhat et al. (2022). Several works build on this approach by utilizing the inherent hierarchical structure of U-Nets to integrate stochasticity Baumgartner et al. (2019); Kohl et al. (2019). Ambiguous data labels can also be exploited by generative models, such as diffusion models Rahman et al. (2023); however, this ambiguity is often not explicitly present in many data sets.

In contrast, generative models for image segmentation such as GANs (Isola et al., 2017; Li et al., 2021) and score-based generative models (Baranchuk et al., 2021; Amit et al., 2021; Wolleb et al., 2022; Wu et al., 2022, 2024) have been utilized to focus on learning the data distributions of segmentation masks given input images. While GANs are based on the principle of learning a direct mapping from noise to data through adversarial training, score-based diffusion models learn the log-gradient (score) of a data distribution by leveraging the idea of SDEs to gradually reverse a noise process. In our previous work, we use a score-based generative model where segmentation masks of corresponding medical images are represented implicitly by using SDFs (Bogensperger et al., 2023). By capturing the underlying data distributions, these models provide more robust predictions, especially in scenarios with complex or unseen data. Although they offer statistical robustness, they generally require more complex training processes and tend to operate slower.

A more recent generative method is flow matching (Lipman et al., 2022; Liu et al., 2022), which extends continuous normalizing flows (CNFs) by a simulation-free training strategy. This method learns a vector field related to a conditional probability path, establishing a connection with diffusion-based methods. Likewise, the goal is to transition from a distribution of interest to a simple, tractable prior distribution - e.g. by gradually injecting noise to smooth a

data distribution. Flow matching has been shown to be a more robust and stable alternative to diffusion models offering enhanced performance (Lipman et al., 2022), and has recently been applied to high-quality, unconditional image synthesis (Esser et al., 2024). Moreover, recent studies demonstrated its effectiveness in various domains such as depth estimation, label-conditioned image generation and image inpainting (Dao et al., 2023; Gui et al., 2024; Yariv et al., 2023). The subsequent section will introduce flow matching in more detail.

2.2 Flow Matching

CNFs (Chen et al., 2018) characterize the target density $q = [\Psi]_{\#} p_0$ as push-forward of a tractable initial density p_0 under a deterministic transformation $\Psi: \mathbb{R}^d \rightarrow \mathbb{R}^d$. Typically, the transformation $\Psi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is learned via maximum likelihood and maps samples $x \sim p_0$ from the initial distribution to samples $\Psi(x) \sim q$ that follow the target (Rezende and Mohamed, 2015). However, CNFs do not directly model Ψ , but rather consider the temporal dynamics of a time-dependent transformation $\Psi_t: [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ with $t \in [0, 1]$. The corresponding neural ODE (Chen et al., 2018) is given as

$$\frac{d}{dt} \Psi_t(x) = v_{\theta,t}(\Psi_t(x)), \quad (1)$$

where $v_{\theta,t}: [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a parameterized network. It defines a time-dependent vector field by assigning spatial displacements to each spatio-temporal pair $(t, \Psi_t(x))$. The overall transformation $\Psi(x) = \Psi_1(x) = x + \int_0^1 v_{\theta,t}(x) dt$ for the initial state x is then given by integrating the spatial displacements over the whole time domain, which can be done by utilizing off-the-shelf ODE solvers.

However, the widespread adoption of CNFs as generative models has so far been hindered by the cost of maximum likelihood training. Because of this limitation, flow matching has been introduced as a feasible alternative (Lipman et al., 2022; Liu et al., 2022). The essential insight is that any time-dependent vector field $u_t: [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ which is related to the temporal dynamics of a given density path $p_t: [0, 1] \times \mathbb{R}^d \rightarrow [0, \infty)$ via the ubiquitous continuity equation (Villani et al., 2009)

$$\frac{d}{dt} p_t(x) + \text{div}(p_t(x) u_t(x)) = 0, \quad (2)$$

can be used as a regression target. The corresponding flow matching objective is thus given as

$$\mathcal{L}_{\text{FM}}(\theta) := \mathbb{E}_{t \sim \mathcal{U}_{[0,1]}} \mathbb{E}_{x \sim p_t} \left[\frac{1}{2} \|v_{\theta,t}(x) - u_t(x)\|^2 \right]. \quad (3)$$

This objective is justified by identifying the density path in Eq. (2) with the one that is induced by the continuous transformation, i.e., $p_t = [\Psi_t]_{\#} p_0$. Thus, it becomes evident that minimizing Eq. (3) indeed yields a vector field $v_{\theta,t}(\cdot)$ characterizing Ψ_t via Eq. (1). However, as the marginal ground truth vector field u_t is not known in general and p_t can take a complicated form, directly minimizing Eq. (3) is impossible. Instead, Lipman et al. (Lipman et al., 2022) show that u_t can be constructed by superimposing many conditional vector fields $u_t(\cdot|x_1)$ that depend on available training samples $x_1 \sim q$. The simpler conditional flow matching objective is thus given as

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}_{[0,1]}} \mathbb{E}_{x_1 \sim q} \mathbb{E}_{x \sim p_t(\cdot|x_1)} \left[\frac{1}{2} \|v_{\theta,t}(x) - u_t(x|x_1)\|^2 \right], \quad (4)$$

which yields the same set of minimizers as Eq. (3) and reveals similarities to score-based optimization (Hyvärinen and Dayan, 2005). To obtain an explicit expression for the regression target $u_t(x|x_1)$, they propose conditional affine Gaussian flows $\Psi_t(x|x_1) = \sigma_t x + \mu_t(x|x_1)$. According to Eq. (1) it is of the form $u_t(x|x_1) = \frac{x - \mu_t(x|x_1)}{\sigma_t} \frac{d}{dt} \sigma_t + \frac{d}{dt} \mu_t(x|x_1)$. In addition, the corresponding density path has the form $p_t(x|x_1) = \mathcal{N}(x|\mu_t(x|x_1), \sigma_t^2 \text{Id})$. Following (Lipman et al., 2022), it starts from the tractable multivariate standard normal distribution $p_0(\cdot|x_1) = \mathcal{N}(\cdot|0, \text{Id})$ and ends in a Dirac peak at the conditioning sample x_1 , i.e., $p_1(\cdot|x_1) = q(\cdot|x_1) = \mathcal{N}(\cdot|x_1, \sigma_{\min}^2 \text{Id})$, $\sigma_{\min}^2 \approx 0$. In contrast, the temporal interpolation between (μ_0, σ_0) and (μ_1, σ_1) is not uniquely defined and offers some degree of freedom. In this work, we adopt the best-performing variant from Lipman et al. (2022) that induces conditional optimal transport paths via $\mu_t(x|x_1) = tx_1$ and $\sigma_t = 1 - (1 - \sigma_{\min})t$.

3 Method

3.1 Image Segmentation Using SDFs

Let M, N be the spatial dimensions of the discretized images with domain $\Omega = \{1, \dots, M\} \times \{1, \dots, N\}$, such that $d := M \times N$. We usually denote an image $x \in \mathcal{X} := \mathbb{R}^{M \times N}$. For C -class semantic image segmentation the images have corresponding segmentation masks $m \in \mathcal{M}$, where $\mathcal{M} := \{0, \dots, C-1\}^{M \times N}$, with $C = 2$ for binary segmentation. Moreover, we consider the SDF representation of a segmentation mask $\tilde{m} \in \tilde{\mathcal{M}}$, which yields a continuous representation such that $\tilde{\mathcal{M}} := \mathbb{R}^{M \times N}$.

Using SDF, the segmentation problem can be reformulated in terms of signed distances relative to the object boundaries. For each pixel m_{ij} in the domain Ω with the object $\mathcal{S}$ to be segmented, the SDF map assigns the distance

to the nearest boundary pixel $\partial\mathcal{S}$. Additionally, the distance is signed, such that a negative sign indicates that the pixel is inside the object, while a positive sign indicates that it lies outside.

The SDF $\tilde{m} \in \tilde{\mathcal{M}}$ of a segmentation mask m can be computed by applying the Euclidian distance function to each pixel m_{ij} . To focus on the region near the object boundary $\partial\mathcal{S}$, the distances can be truncated at a threshold δ . This truncation also prevents distant background pixels from influencing the segmentation outcome, as large positive distances are ignored. The implicit, truncated SDF $\phi(m_{ij})$ is then given by (also see (Osher et al., 2004)):

$$\phi(m_{ij}) = \begin{cases} -\min\{\min_{y \in \partial\mathcal{S}} \|y - m_{ij}\|_2, \delta\} & \text{if } m_{ij} \in \mathcal{S}, \\ \min\{\min_{y \in \partial\mathcal{S}} \|y - m_{ij}\|_2, \delta\} & \text{if } m_{ij} \in \Omega \setminus \mathcal{S}, \\ 0 & \text{if } m_{ij} \in \partial\mathcal{S}. \end{cases} \quad (5)$$

Thus we can obtain the full segmentation map $\tilde{m} = \phi(m_{ij})_{i=1, \dots, M, j=1, \dots, N}$ by applying $\phi(\cdot)$ on the entire segmentation mask m , which is also depicted in Fig. 2. Conversely, given $\tilde{m}$ one can easily retrieve the binary segmentation map m by thresholding at 0 to separate segmented objects from background, as given by $m = \mathbf{1}_{\{\tilde{m} \leq 0\}}$, where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function that evaluates to 1 or 0 depending on whether the respective condition is fulfilled.

3.2 Motivating the SDF

In a naive approach, the push forward function $p_t = [\Psi_t]_{\#} p_0$ transitions the distribution of binary masks towards a tractable multivariate Gaussian by incrementally adding noise to the desired segmentation mask. However, the intermediate masks $m_t \sim p_t(m|m_1)$ resulting from this process suffer from arbitrary pixel-wise perforations. This degradation occurs due to the discontinuous nature of segmentation masks. If additive noise causes the decision boundary to be surpassed, i.e. leading to a class jump, it results in arbitrary holes in the mask rather than the gradual removal of semantic information.

By transforming discontinuous masks using distance transforms, a smoother representation is provided while keeping the semantic information. In this setting the perturbation affects the masks starting from boundaries, subsequently progressing towards the center of objects until converging towards a sample from the tractable prior distribution. Therefore using SDF representations of semantic masks introduces a more canonical way of converting the distribution of interest to the tractable prior. A comparison of the resulting transition process on both variants of segmentation mask representations for several intermediate time steps is shown in Fig. 3.

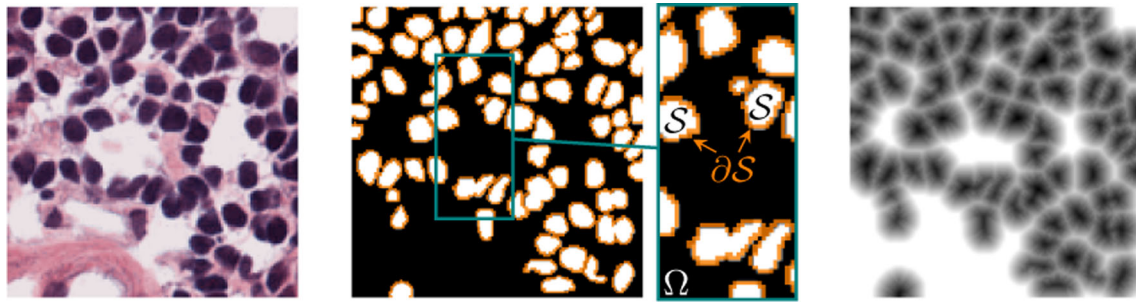


Fig. 2 Given an image x (left), its binary segmentation mask (center) can be transformed into a truncated SDF mask (right). The zoomed area shows some segmentation objects in detail denoted by S and their boundaries ∂S embedded in the domain Ω

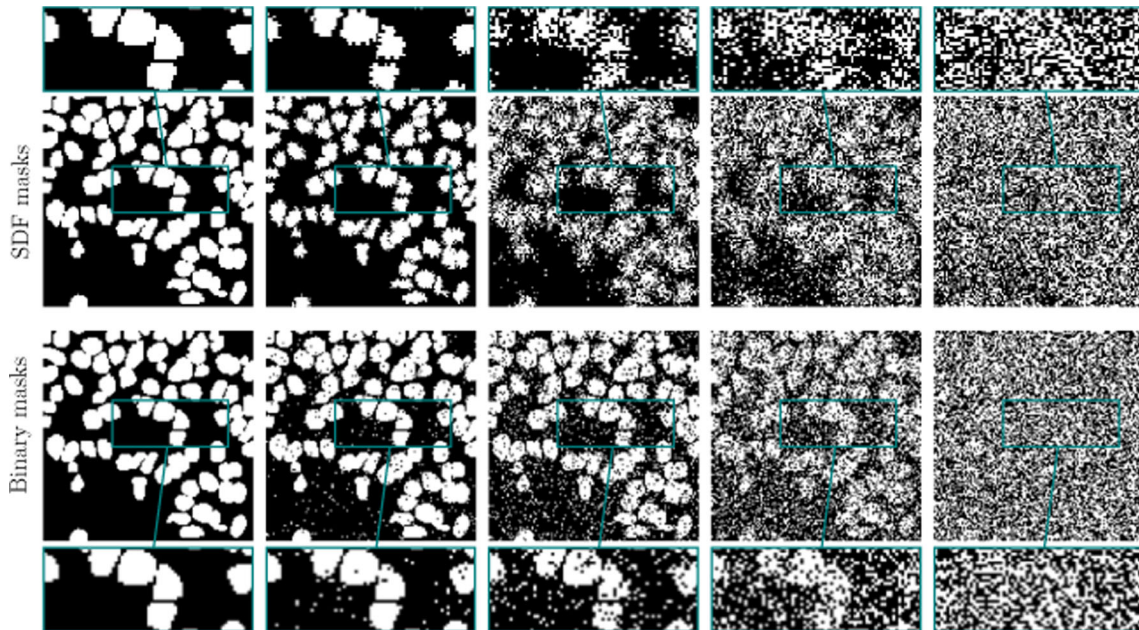


Fig. 3 Comparison of the effect of the transitioning process on the resulting thresholded segmentation masks when using an SDF mask $\tilde{m}$ and a binary mask m for a given image x . Note that the SDF representation allows for a more natural distortion process in its thresholded

masks, which evolves along the object boundaries instead of directly introducing “hole”-like structures at random pixel positions as it is the case in the thresholded masks when directly using the binary segmentation mask

3.3 Image-Guided Flow Matching

Given a data set of S labeled pairs of images and segmentation masks $\mathcal{D} = \{(\tilde{m}_s, x_s)\}_{s=1}^S$, the goal is to learn the conditional distribution of SDF-represented segmentation masks given their respective images $q(\tilde{m}|x)$. Building on the introduced concept of flow matching from Sect. 2.2, the aim is to learn a conditional vector field $v_{\theta,t}(\tilde{m}|x)$ which is related to the sought conditional flow by (cf. Equation (1))

$$\frac{d}{dt}\Psi_t(\tilde{m}|x) = v_{\theta,t}(\Psi_t(\tilde{m}|x), x). \quad (6)$$

This is also depicted graphically in Fig. 4, showing how the conditioning image x guides the flow from the clean SDF mask $\tilde{m}_1$ to a corrupted version $\tilde{m}_0$, which can be reversed by numerically integrating the ODE. Moreover, the

corresponding thresholded binary segmentation masks m_t during the different stages can be observed in the bottom row, however, note that they are only indirectly obtained while the image-guided flow matching explicitly works on the SDF-transformed masks $\tilde{m}_t$. Then the final objective that is minimized is given by

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}_{[0,1]}} \mathbb{E}_{(\tilde{m}_1, x) \sim q(\tilde{m}|x)} \mathbb{E}_{\tilde{m}_t \sim p_t(\cdot|\tilde{m}_1, x)} \left[\frac{1}{2} \|v_{\theta,t}(\tilde{m}_t, x) - u_t(\tilde{m}|\tilde{m}_1, x)\|^2 \right]. \quad (7)$$

Learning the parameters θ of a neural network $v_{\theta,\cdot}$ such that it minimizes Eq. (7) is described in Algorithm 1. Note that $v_{\theta,\cdot}(\cdot, x)$ additionally requires the conditioning image x as an input. For the prior distribution $p(\tilde{m}_0)$ a standard normal distribution $\mathcal{N}(0, \text{Id})$ is used, although flow matching would also allow other choices.

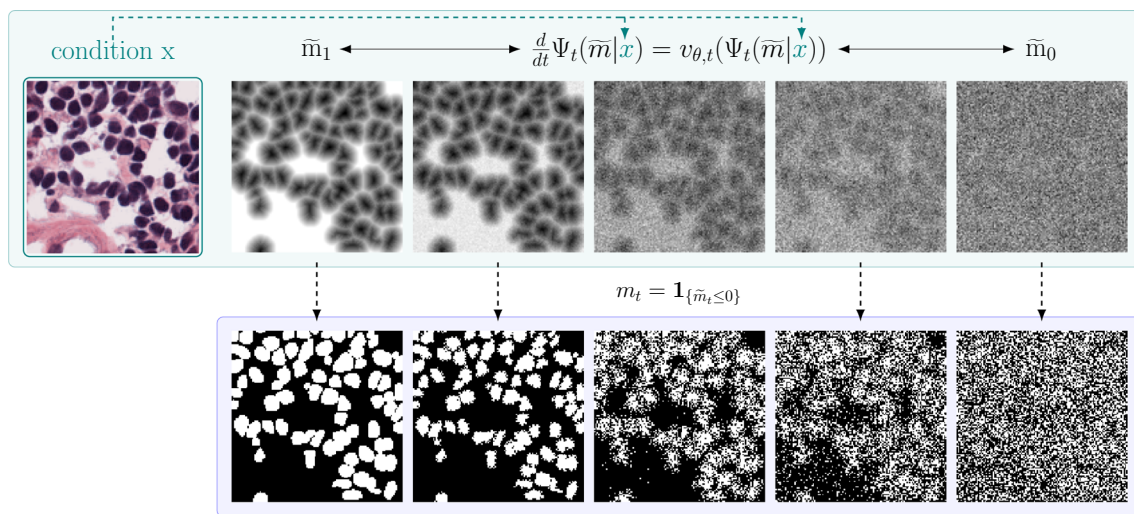


Fig. 4 Schematic of the corruption/generative process (top row for different $t \in [0, 1]$) on an SDF segmentation mask $\tilde{m}$ for a given image x . Note that the governing mechanism is bidirectional, hence to obtain samples during inference a numerical ODE integrator is used. The cor-

responding binary segmentation masks m_t are shown in the second row. While they are not directly used neither during training nor sampling, they can be obtained easily by thresholding and they demonstrate the induced corruption mechanism through the usage of the SDF

Algorithm 1 Image-guided flow matching training

Input data set $\mathcal{D} = \{(\tilde{m}_s, x_s)\}_{s=1}^S \sim q(\tilde{m}|x)$
repeat
 Sample $(\tilde{m}_1, x) \sim \mathcal{D}$
 Sample $t \sim \mathcal{U}_{[0,1]}$
 Sample $\tilde{m}_0 \sim p(\tilde{m}_0)$
 $\tilde{m}_t \leftarrow t\tilde{m}_1 + (1-t)\tilde{m}_0$
 $\tilde{m}_t \leftarrow \tilde{m}_1 - \tilde{m}_0$
 $\theta^{k+1} = \text{Adam}(\theta^k, \nabla_{\theta^k} \frac{1}{2} \|v_{\theta^k}(\tilde{m}_t, x) - \tilde{m}_t\|_2^2)$
until convergence

Algorithm 2 Image-guided sampling

trained $v_{\theta,\cdot}$, conditioning image x , ODE discretization $\eta = \frac{1}{\text{NFE}}$,
 noise injection steps $T \in \mathbb{N}$ with $\tau = \frac{1}{T}$
 $\tilde{m}_0 \sim p(\tilde{m}_0)$
for $t = 0, \dots, 1 - \tau$ **do**
for $l = t, \dots, 1 - \eta$ **do**
 $\tilde{m}_t^{l+\eta} \leftarrow \text{ODEstep}(v_{\theta,l}(\tilde{m}_t^l, x))$
end for
 $z \sim \mathcal{N}(0, I)$
 $\tilde{m}_{t+\tau} \leftarrow (t + \tau)\tilde{m}_t^1 + (1 - t - \tau)z$
end for
 $\tilde{m}_1 \leftarrow \tilde{m}_t$
return $\tilde{m}_1$

Generating new samples using flow matching involves integrating the ODE in Eq. (6), the procedure is given in detail in Algorithm 2. In general, one has to choose the number of function evaluations (NFE) which corresponds to the discretization of the ODE with η . Moreover, to compare with sampling based on SDEs, T noise injection steps can be incorporated to reintroduce noise into the denoised mask $\tilde{m}_1$.

with a linear schedule. If $T = 1$, the original ODE-based sampling as given in (Lipman et al., 2022) is recovered.

By employing a generative model, the goal shifts from learning a discriminative model that predicts segmentation masks in an end-to-end manner to a model that ideally captures the underlying statistics of the conditional distribution $q(\tilde{m}|x)$. As the noise injected inference procedure resembles a stochastic sampling process, individual runs can be seen as samples $\tilde{m}_k \sim p(\tilde{m}|x)$. Therefore, the minimum mean square error (MMSE) can be approximated by $\mathbb{E}[\tilde{m}|x] \approx \frac{1}{K} \sum_{k=1}^K \tilde{m}_k$ with K independent inference runs. Furthermore, the variance $\mathbb{V}[\tilde{m}|x] \approx \frac{1}{K} \sum_{k=1}^K (\tilde{m}_k - \mathbb{E}[\tilde{m}|x])^2$ can be approximated and yields insightful information about classification uncertainty.

4 Experiments

4.1 Data Sets and Preprocessing

For the experimental setting, we utilize two publicly available data sets. The first data set is MoNuSeg (Kumar et al., 2017), which consists of 30 training images and 14 test images. Each image is of size 1000×1000 and overall they contain more than 21,000 annotated nuclei in H&E stained microscopic images. For data preprocessing we resort to a structure-preserving color normalization (Vahadane et al., 2015) as the different organ sites yield considerable intensity variations in the data. Subsequently, all images were resized to 500×500 following (Valanarasu et al., 2021). During training, random crops of 128×128 were used with additional random horizontal and vertical flips for data augmentation

As a second medical data set, the Gland Segmentation (GlaS) data set (Sirinukunwattana et al., 2017) was used. It consists of 85 training and 80 test H&E stained microscopic images with annotated glands from colorectal cancer tissue. Again, a structure-preserving color normalization (Vahadane et al., 2015) is used also here for data preprocessing, which is followed by resizing all training and test images to 128×128 inspired by (Valanarasu et al., 2021). Data augmentation in training is done as with the MoNuSeg data set.

4.2 Training Details

The deep learning architecture to approximate the vector field $v_{\theta, \cdot}(\cdot, x)$ is based on work presented in (Nichol and Dhariwal, 2021; Amit et al., 2021). The mechanism to condition the learned vector field on the image x are adopted from (Amit et al., 2021), where the extracted features of the noisy mask $\tilde{m}_t$ are additively combined with the feature maps of the conditioning image x . Note that this differs from other commonly used conditioning mechanisms that concatenate the extracted features (Özdenizci and Legenstein, 2023; Ho et al., 2022; Saharia et al., 2022), but the extensive ablation study over the conditioning mechanism in (Amit et al., 2021) shows superior performance.

During training, we utilized the Adam optimizer with a learning rate of $1e-4$. The batch size was set to 16, and the models were trained for approximately 30k epochs for the first data set and nearly 40k epochs for the second. The average training time per epoch was 5.4 s for the MoNuSeg data set and 15.5 s for the GlaS data set, using an NVIDIA Titan RTX GPU. Additionally, (EMA) with a decay factor of 0.9999 was applied throughout the training process. The detailed training procedure of image-guided flow matching is described in Algorithm 1.

4.3 Evaluation

After training we obtain a network $v_{\theta, \cdot}$ that regresses a vector field, capable of gradually transitioning $p(\tilde{m}_0)$ to $p(\tilde{m}_1|x)$. Consequently, Algorithm 2 can be employed to synthesize new samples $\tilde{m}_1$ given the conditioning images x that have to be segmented. The ODEstep is implemented in practice using numerical integration provided by (Chen, 2018), requiring about 0.85 s to generate 4 samples with 4 NFEs. The number of NFE is set to 4, as flow matching in general already yields remarkable results with only few NFEs, which is advantageous compared to standard DDPMs. Moreover, averaging over a number of K predictions to obtain an estimate for the MMSE was considered, where the respective ablation that investigates the choice of K can be found in Fig. 9 for both data sets. For multiple noise injection runs T , we set $T = 100$ to enable a fair comparison with the SDE sampler in (Amit et al., 2021).

As the image size was set to 128×128 to reduce computational demands, the inference procedure for the MoNuSeg data set consists of overlapping patches of each test image. Note that for GlaS this was not necessary, as the entire images were resized to 128×128 . Hence, an averaging strategy, similar to that proposed by (Özdenizci and Legenstein, 2023), was used to merge the overlapping sections within each image. Thereby, the average of the predictions over 2–4 different patch segments was taken to diminish the effect of cropped nuclei at the borders. This also quantitatively improved the evaluation results further, see Table 3.

The obtained segmentation masks are then evaluated using the standard metrics F1 score and Intersection over Union (IoU). We compare our method to commonly referred benchmark models, ensuring that both U-Net variants and attention/transformer mechanisms are considered. Moreover, we also consider (Amit et al., 2021) using the publicly available source code to obtain a comparison with a generative, conditional DDPM that predict standard binary segmentation masks in the sampling process. Finally, the approach in (Bogensperger et al., 2023) is also compared with, which is a conditional diffusion model based on a variance-exploding SDE where also SDF masks are predicted in the sampling process.

4.4 Results

Table 1 shows quantitative results for both data sets with our method and comparison methods. To enable a fair comparison the benchmark results were taken from (Valanarasu et al., 2021) where possible, apart from SegDiff (Amit et al., 2021). If not specified, the number of additional noise injection steps $T = 1$ is used with FlowSDF. The results indicate that our method reaches competitive results on both data sets, already with a single inference run with $K = 1$. Additionally, for both data sets using the MMSE by averaging over multiple K sampling runs gives a significant boost in quantitative performance. Moreover, using $T = 100$ noise injection steps as described in Algorithm 2 further improves the quantitative metrics, yielding state-of-the-art results.

Note that for ODE-based sampling with FlowSDF (i.e. $K = 1$), the predictions for a single run seem to be closer to the MMSE than with SDE-based sampling. Hence, diffusion models experience a greater boost by increasing K as observed in (Amit et al., 2021), whereas FlowSDF already exhibits high-quality results with $K = 1$.

Figure 5 shows exemplary qualitative results for both data sets to further highlight the potential applicability of FlowSDF. The segmented objects m derived from thresholding the SDF predictions $\tilde{m}_1$ are distinctly cohesive and closely match the ground truth segmentation masks m_{gt} for both data set samples.

Table 1 Quantitative segmentation results on the MoNuSeg and GlaS data set.

Method	MoNuSeg F1 $\uparrow$	mIoU $\uparrow$	GlaS F1 $\uparrow$	mIoU $\uparrow$
FCN (Badrinarayanan et al., 2017)	28.84	28.71	66.61	50.84
U-Net (Ronneberger et al., 2015)	79.43	65.99	77.78	65.34
U-Net++ (Zhou et al., 2018)	79.49	66.04	78.03	65.55
Res-UNet (Xiao et al., 2018)	79.49	66.07	78.83	65.95
Axial Attention U-Net (Wang et al., 2020)	76.83	62.49	76.26	63.03
MedT (Valanarasu et al., 2021)	79.55	66.17	81.02	69.61
SegDiff (Amit et al., 2021) (K=1)	76.88	62.59	86.40	77.36
SegDiff (Amit et al., 2021) – MMSE (K=25)	81.19	68.44	87.27	78.65
VE SDE with SDF (Bogensperger et al., 2023)	78.13	64.19	82.03	71.36
FlowSDF (K=1)	80.67	67.70	88.64	80.63
FlowSDF – MMSE (K=25)	81.02	68.19	<u>88.94</u>	<u>81.05</u>
FlowSDF – MMSE (K=25, T=100)	81.22	68.46	89.31	81.71

Best score for each metric in **bold**, second-best underlined

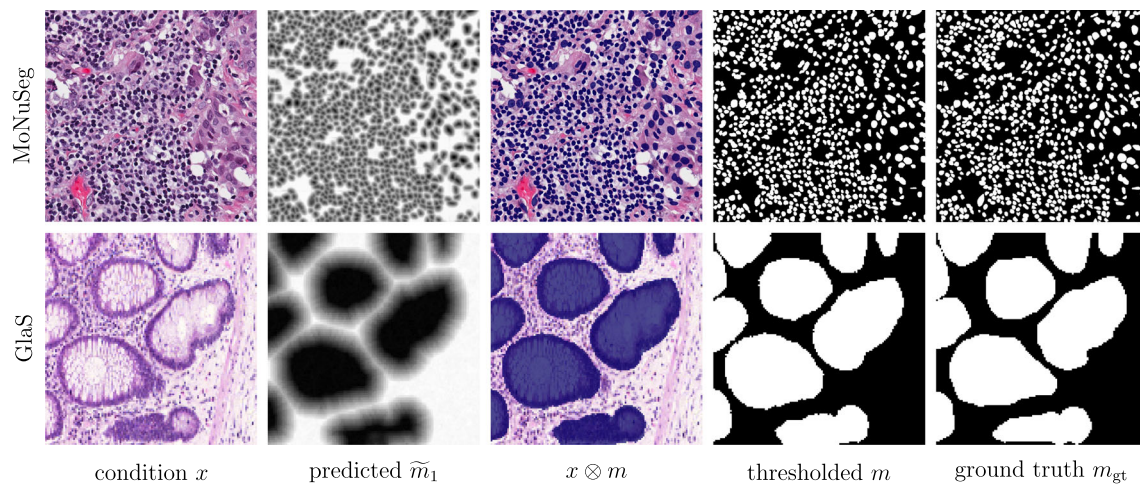


Fig. 5 Exemplary sampled segmentation masks for both data sets. The predicted SDF masks $\tilde{m}_1$ are directly obtained from the OODE sampling, whereas the thresholded masks m are shown to additionally

enable a visual comparison with the depicted ground truth m_{gt} . Furthermore, we provide the condition image with the overlaid thresholded mask $x \otimes m$

A visual comparison is additionally depicted in Fig. 6, where we compare the thresholded segmentation prediction of FlowSDF for a MoNuSeg test image with the DDPM-based (generative) model (Amit et al., 2021) and two discriminative models, namely U-Net++ (Zhou et al., 2018) and MedT (Valanarasu et al., 2021). As can be seen in the provided zoom, the SDF representation of segmentation objects indeed seems to act as a shape prior and thus yields smoother segmentation objects while avoiding artefacts such as single pixels/small structures that are mistakenly classified as foreground objects. Note that SegDiff requires more runs ($K = 25$ as reported in (Amit et al., 2021)) in order to increase performance, while FlowSDF already yields comparatively good results with one evaluation cycle ($K = 1$),

likely due to a combination of the SDF representation and the ODE-based sampling.

Utilizing a generative model enables us to sample from the conditional distribution of the SDF given the conditioning image x . An advantage of this approach is given by the fact that the resulting statistical values such as the pixel-wise variance $\mathbb{V}[\tilde{m}|x]$ allow for the quantification of segmentation uncertainties in the SDF predictions as well as the thresholded masks, which provide additional insights into the segmentation process that are not available with traditional discriminative point estimators. An illustration of the aforementioned property on a MoNuSeg test image x can be seen in Fig. 7 for $K = 128$, where the standard deviation maps associated with the SDF predictions and thresholded masks highlight the regions of uncertainty.

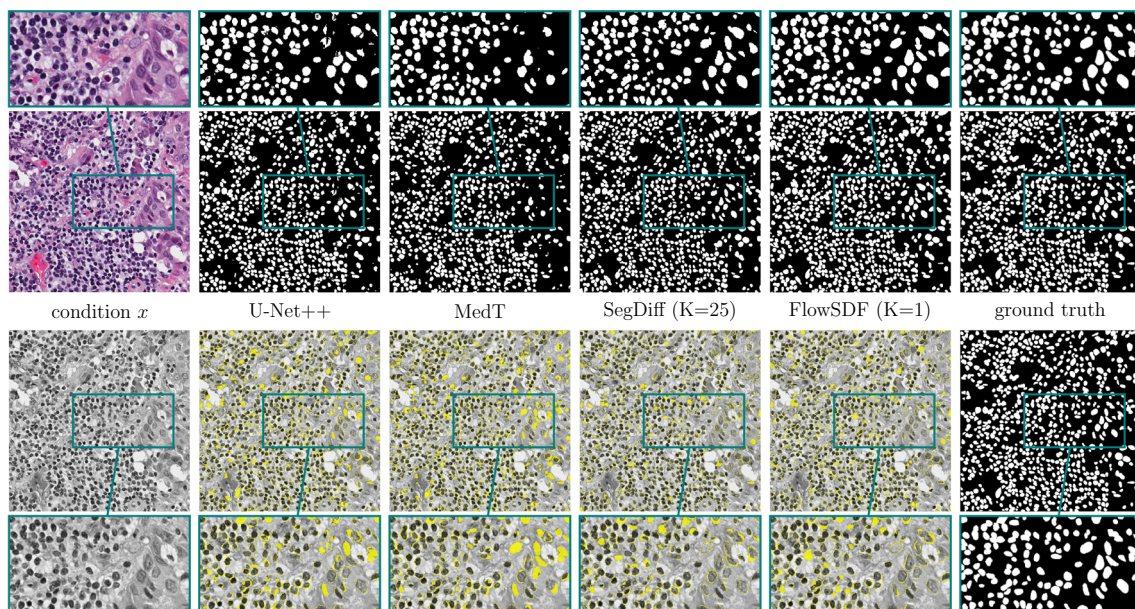


Fig. 6 Qualitative results on a MoNuSeg test image x for different selected discriminative and generative approaches. Top row: we compare FlowSDF to the generative DDPM SegDiff (Amit et al., 2021) and two discriminative models, U-Net++ (Zhou et al., 2018) and MedT (Valanarasu et al., 2021). Bottom row: we present the respec-

tive difference images to the ground truth blended into the gray-scaled condition image for the approaches reported above. As depicted in the provided zoom, the generative methods yield smoother objects, which can be attributed to averaging over $K = 25$ runs for Segdiff and to the SDF with FlowSDF, where $K = 1$ suffices

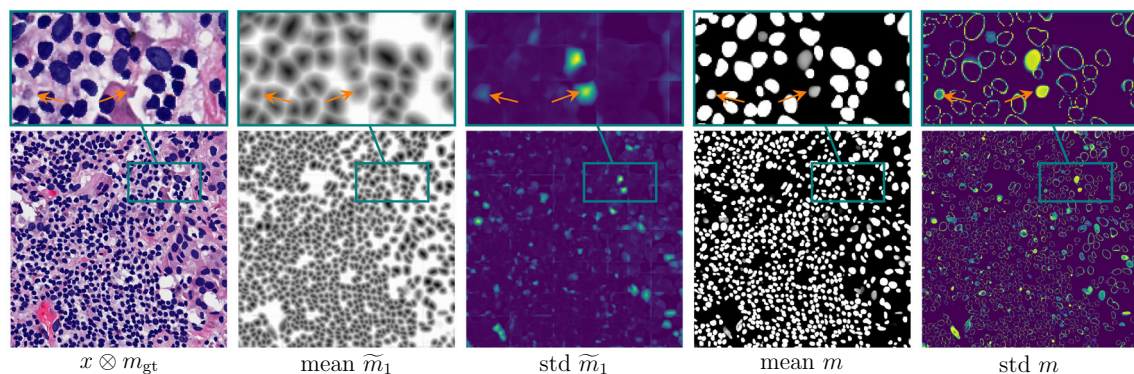


Fig. 7 Segmentation example with according statistical values for the SDF predictions $\tilde{m}_1$ respectively thresholded masks m ($K=128$). Notably, the image includes a region where erroneously two nuclei

were segmented, as indicated by the orange arrows. This region is highlighted in the standard deviation maps, which represent the associated uncertainty in the segmentation

In general, we observe that the standard deviation appears high on transitions from nuclei to background as well as in wrongly detected nuclei or over-segmented parts of nuclei. This encouraging observation leads us to the hypothesis that the uncertainty may be associated directly with segmentation errors similar to what has been shown in (Narnhofer et al., 2022), see also Fig. 8. Here, a visual comparison of the error and standard deviation (uncertainty) maps for the same test sample as in Fig. 7 indicates that the latter very likely has high predictive capability. A detailed analysis including the

mutual information of the two variables, however, is out of scope for this work.

4.5 Ablation

We consider several ablation studies, namely the effect of using SDF representations for segmentation masks, the patch-averaging strategy to combine overlapping segments in the MoNuSeg test data and the number of ensemble runs K when computing MMSE predictions. Table 2 compares

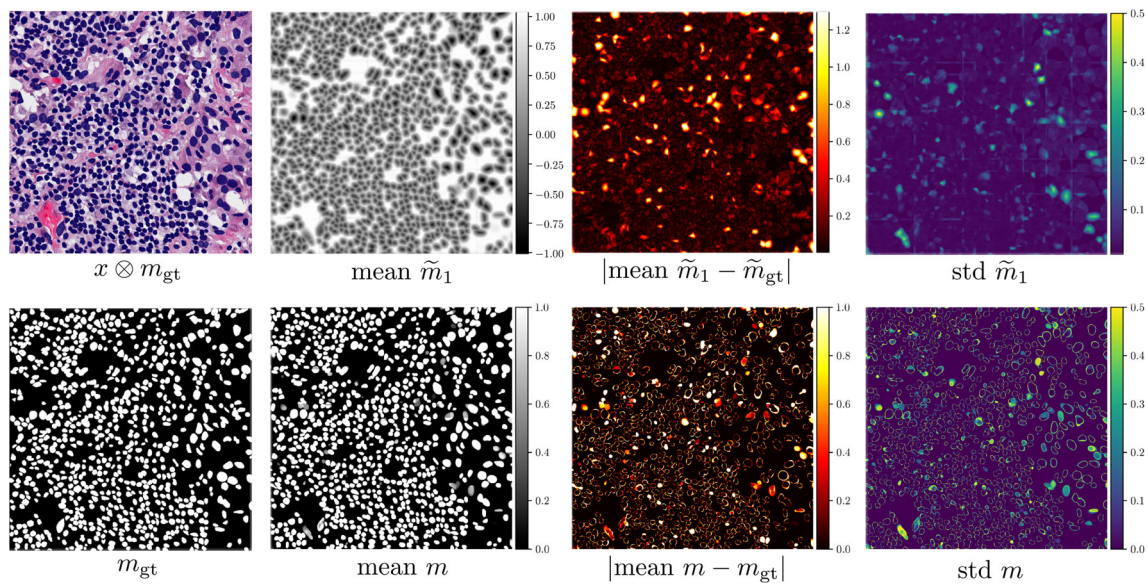


Fig. 8 Segmentation example with statistical values for the predicted $\tilde{m}_1$ and thresholded masks m , along with the corresponding ground truth images and the absolute error between the predictions and the ground

truth ($K=128$). The visual comparison of error and standard deviation maps (uncertainty) suggests that the latter has predictive capability for the error

Table 2 Ablation results over SDF representation compared to binary mask

Method	MoNuSeg F1 $\uparrow$	mIoU $\uparrow$	GlaS F1 $\uparrow$	mIoU $\uparrow$
w/o SDF	79.78	66.45	88.20	79.91
w SDF	80.67	67.70	88.64	80.63

Table 3 Influence of patch-averaging strategy for overlapping segments in MoNuSeg

Method	MoNuSeg F1 $\uparrow$	mIoU $\uparrow$
w/o patch-averaging	80.45	67.38
w patch-averaging	80.67	67.70

quantitative segmentation results for both data sets using binary masks (w/o SDF) and FlowSDF (w SDF), showing the performance increase due to using SDFs. Note that the results are obtained using a single run with $K = 1$, but the same trend is also observed using ensemble predictions.

Moreover, for evaluating test images of the MoNuSeg data set, overlapping patches of 128×128 as described in Sect. 4.3 were extracted. The effect of averaging over its merged regions similar to (Özdenizci and Legenstein, 2023) is shown in Table 3. Note that this is shown only for the MoNuSeg data set where the available images are too large to directly resize the whole images, as opposed to the GlaS data set which allowed for evaluations on the entire images.

Finally, the effect of K independent sampling runs as described in Algorithm 2 to approximate the MMSE is visu-

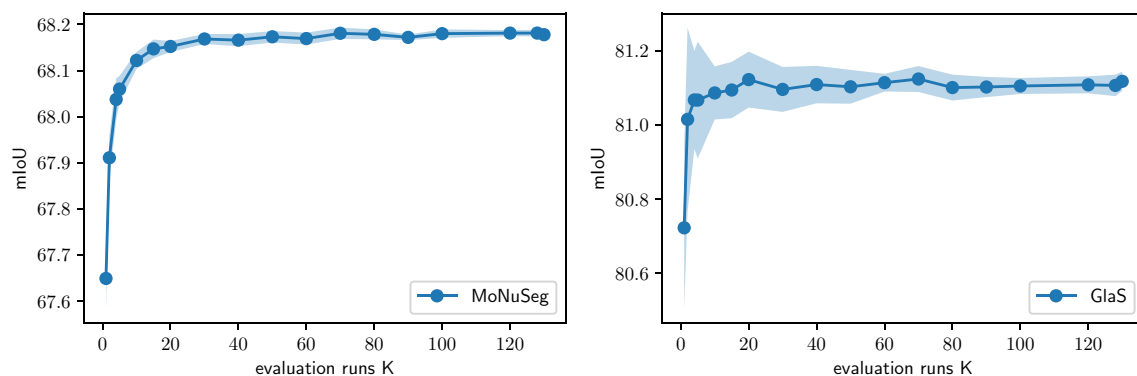


Fig. 9 Influence of evaluation runs K on mIoU averaged over 25 runs for both data sets. Note that ODE-based sampling initially benefits from increased runs K but soon stabilizes and remains fairly constant

alized in Fig. 9 for both data sets. We use only one noise injection step T here, to recover the original ODE-based sampler as described in (Lipman et al., 2022). It is noteworthy that unlike SDE-based samplers that are used in diffusion models, ODE samplers seem to better approximate the MMSE already for smaller values of K . Thus, the performance gain is not as large although increasing K moderately definitely helps in boosting quantitative metrics.

5 Conclusion

In this work we proposed FlowSDF, a generative approach for medical image segmentation by fusing image-guided flow matching with the concept of representing segmentation maps with the SDF. The applicability of FlowSDF was demonstrated qualitatively and quantitatively on two public medical data sets. The SDF provides a smoother and more continuous representation of object boundaries compared to binary masks, which makes it a more canonical choice for accurate and robust segmentation. Moreover, by leveraging the generative approach, statistical measures such as mean and standard deviation can be computed to quantify the uncertainty in the segmentation results. This information is especially useful for medical diagnosis and treatment planning, where it is important to know the level of confidence in the segmentation results.

In the future, it is of interest to investigate the benefit of the SDF in object detection and segmentation beyond the medical domain where data sets contain more complex shapes that are in general non-circular. The extension to multi-class semantic segmentation could also be studied. Finally, with the success of latent diffusion models (LDMs) (Rombach et al., 2022), the usage of foundation models should be investigated. As LDMs have also been developed in the medical domain (Yellapragada et al., 2024), it is of high interest to study the extent to which they can capture semantic concepts and potentially improve upon existing methods. With the outlook of a recent, powerful foundational model that is based on flow matching (Esser et al., 2024), their usage in a conditional setting for image segmentation can be explored. Such approaches have already been successfully applied for learning conditional distributions in depth estimation (Gui et al., 2024; Ke et al., 2024).

Funding Open access funding provided by Graz University of Technology.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence,

unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amit, T., Nachmani, E., Shaharbany, T., Wolf, L. (2021). Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint [arXiv:2112.00390](https://arxiv.org/abs/2112.00390)
- Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.
- Baranchuk, D., Rubachev, I., Voynov, A., Khruikov, V., Babenko, A. (2021). Label-efficient semantic segmentation with diffusion models. arXiv preprint [arXiv:2112.03126](https://arxiv.org/abs/2112.03126)
- Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E. (2019). Phiseg: Capturing uncertainty in medical image segmentation. In Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, 119–127. Springer.
- Bhat, I., Pluim, J. P. W., & Kuijff, H. J. (2022). Generalized Probabilistic U-Net for Medical Image Segmentation. In C. H. Sudre, C. F. Baumgartner, A. Dalca, C. Qin, R. Tanno, K. Van Leemput, & W. M. Wells III (Eds.), *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings* (pp. 113–124). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16749-2_11
- Bogensperger, L., Narnhofer, D., Ilic, F., & Pock, T. (2024). Score-Based Generative Models for Medical Image Segmentation Using Signed Distance Functions. In U. Köthe & C. Rother (Eds.), *Pattern Recognition: 45th DAGM German Conference, DAGM GCPR 2023, Heidelberg, Germany, September 19–22, 2023, Proceedings* (pp. 3–17). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54605-1_1
- Boykov, Y., & Funka-Lea, G. (2006). Graph cuts and efficient nd image segmentation. *International Journal of Computer Vision*, 70(2), 109–131.
- Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), 1124–1137.
- Brissman, E., Johnander, J., Felsberg, M. (2021). Predicting signed distance functions for visual instance segmentation. In 2021 Swedish Artificial Intelligence Society Workshop (SAIS), pp. 1–6. IEEE
- Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023). Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* (pp. 205–218). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25066-8_9
- Chan, T. F., & Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on Image Processing*, 10(2), 266–277.
- Chen, R.T.Q. (2018). torchdiffeq. <https://github.com/rtqichen/torchdiffeq>

- Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31
- Dao, Q., Phung, H., Nguyen, B., Tran, A. (2023). Flow matching in latent space. *arXiv preprint* [arXiv:2307.08698](https://arxiv.org/abs/2307.08698)
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint* [arXiv:2403.03206](https://arxiv.org/abs/2403.03206)
- Gui, M., Fischer, J.S., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B. (2024). Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint* [arXiv:2403.13788](https://arxiv.org/abs/2403.13788)
- Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., & Salimans, T. (2022). Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47), 1–33.
- Hyvärinen, A., Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).
- Iqbal, A., Sharif, M., Yasmin, M., Raza, M., & Aftab, S. (2022). Generative adversarial networks and its applications in the biomedical image segmentation: a comprehensive survey. *International Journal of Multimedia Information Retrieval*, 11(3), 333–368.
- Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.
- Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: active contour models. *International Journal of Computer Vision*, 1(4), 321–331.
- Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K. (2024). Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O. (2019). A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint* [arXiv:1905.13077](https://arxiv.org/abs/1905.13077)
- Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O. (2018). A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31.
- Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., & Sethi, A. (2017). A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7), 1550–1560.
- Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S. (2021). Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8300–8311.
- Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M. (2022). Flow matching for generative modeling. *arXiv preprint* [arXiv:2210.02747](https://arxiv.org/abs/2210.02747)
- Liu, X., Gong, C. (2022). Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*
- Mumford, D., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42(5), 577–685. <https://doi.org/10.1002/cpa.3160420503>
- Narnhofer, D., Habring, A., Holler, M., Pock, T. (2022). Posterior-variance-based error quantification for inverse problems in imaging. *ArXiv* [arXiv:2212.12499](https://arxiv.org/abs/2212.12499)
- Naylor, P., Laé, M., Reyat, F., & Walter, T. (2018). Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38(2), 448–459.
- Nichol, A.Q., Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.
- Osher, S., & Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79(1), 12–49.
- Osher, S., Fedkiw, R., & Piechor, K. (2004). Level set methods and dynamic implicit surfaces. *Applied Mechanics Reviews*, 57(3), 15–15.
- Özdenizci, O., & Legenstein, R. (2023). Restoring Vision in Adverse Weather Conditions With Patch-Based Denoising Diffusion Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 10346–10357. <https://doi.org/10.1109/TPAMI.2023.3238179>
- Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M. (2023). Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11536–11546
- Rezende, D., Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, 1530–1538.
- Rezende, D., Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, 1530–1538.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695
- Ronneberger, O., Fischer, P., Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18, pp. 234–241 . Springer
- Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M. (2022). Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.
- Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., & Sanchez, U. (2017). Gland segmentation in colon histology images: the glas challenge contest. *Medical Image Analysis*, 35, 489–502.
- Vahadane, A., Peng, T., Albarqouni, S., Baust, M., Steiger, K., Schlitter, A.M., Sethi, A., Esposito, I., Navab, N. (2015). Structure-preserved color normalization for histological images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 1012–1015. IEEE.
- Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24, 36–46. Springer
- Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C. (2020) Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, Proceedings, Part IV*, pp. 108–126 (2020). Springer
- Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C. (2022). Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, 1336–1348. PMLR.
- Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y. (2022). Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint* [arXiv:2211.00611](https://arxiv.org/abs/2211.00611)

- Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., & Xu, Y. (2024). Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, 6030–6038.
- Xiao, X., Lian, S., Luo, Z., Li, S. (2018). Weighted res-unet for high-quality retina vessel segmentation. In 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 327–331. IEEE.
- Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W., & Huang, X. (2020). Shape-Aware Organ Segmentation by Predicting Signed Distance Maps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 12565–12572. <https://doi.org/10.1609/aaai.v34i07.6946>
- Xun, S., Li, D., Zhu, H., Chen, M., Wang, J., Li, J., Chen, M., Wu, B., Zhang, H., & Chai, X. (2022). Generative adversarial networks in medical image segmentation: a review. *Computers in Biology and Medicine*, 140, 105063.
- Yariv, L., Puny, O., Neverova, N., Gafni, O., Lipman, Y. (2023). Mosaic-sdf for 3d generative models. arXiv preprint [arXiv:2312.09222](https://arxiv.org/abs/2312.09222)
- Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D. (2024). Pathldm: Text conditioned latent diffusion model for histopathology. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 5182–5191
- Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., & Li, H. (2023). Diffusion-based data augmentation for nuclei image segmentation. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, & R. Taylor (Eds.), *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023* (pp. 592–602). Cham: Springer.
- Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, 3–11. Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.