MELBA Journal Machine Learning for Biomedical Imaging

ContourDiff: Unpaired Medical Image Translation with Structural Consistency

Yuwen **Chen** [1], Nicholas **Konz** [1], Hanxue **Gu** [1], Haoyu **Dong** [1], Yaqian **Chen** [1], Lin **Li** [2], Jisoo **Lee** [3], Maciej A. **Mazurowski** [1,2,3,4]

1 Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708
2 Department of Biostatistics & Bioinformatics, Duke University, Durham, NC 27708
3 Department of Radiology, Duke University, Durham, NC 27708
4 Department of Computer Science, Duke University, Durham, NC, 27708

Abstract

Accurately translating medical images between different modalities, such as Computed Tomography (CT) to Magnetic Resonance Imaging (MRI), has numerous downstream clinical and machine learning applications. While several methods have been proposed to achieve this, they often prioritize perceptual quality with respect to output domain features over preserving anatomical fidelity. However, maintaining anatomy during translation is essential for many tasks, e.g., when leveraging masks from the input domain to develop a segmentation model with images translated to the output domain. To address these challenges, we propose ContourDiff with Spatially Coherent Guided Diffusion (SCGD), a novel framework that leverages domain-invariant anatomical contour representations of images. These representations are simple to extract from images, yet form precise spatial constraints on their anatomical content. We introduce a diffusion model that converts contour representations of images from arbitrary input domains into images in the output domain of interest. By applying the contour as a constraint at every diffusion sampling step, we ensure the preservation of anatomical content. We evaluate our method on challenging lumbar spine and hip-and-thigh CT-to-MRI translation tasks, via (1) the performance of segmentation models trained on translated images applied to real MRIs, and (2) the foreground FID and KID of translated images with respect to real MRIs. Our method outperforms other unpaired image translation methods by a significant margin across almost all metrics and scenarios. Moreover, it achieves this without the need to access any input domain information during training and we further verify its zero-shot capability, showing that a model trained on one anatomical region can be directly applied to unseen regions without retraining. Our code is available at `https://github.com/mazurowski-lab/ContourDiff`.

Keywords

Unpaired Image Translation, Medical Image Segmentation, Diffusion Model

Article informations

`https://doi.org/https://doi.org/10.59275/j.melba.2025-79a2`

Volume 3, Received: 2025-07-10, Published 2025-11-25

Corresponding author: yuwen.chen@duke.edu

1. Introduction

Unpaired image-to-image (I2I) translation—the task of translating images from some input domains to an output domain with only unpaired data for training Zhu et al. (2017)—offers extensive applications in medical image analysis Armanious et al. (2020); Durrer et al. (2024); Beizaee et al. (2023); Wang et al. (2024); Modanwal et al. (2020); Yang et al. (2019); Liu et al. (2021); Zhang et al. (2018). A significant use case is facilitating segmentation across different imaging modalities (e.g., CT and MRI) Chen et al. (2023a), for anatomical locations such as brain Li et al. (2020), abdomen Huo et al. (2019), and pelvis Rossi and Cerveri (2021). This approach is especially beneficial given the significant time and labor involved in annotating images for each modality independently. Through direct image translation between modalities, annotations from one modality can be reused in another, reducing manual effort. However, achieving this requires strict anatomical consistency in translation.

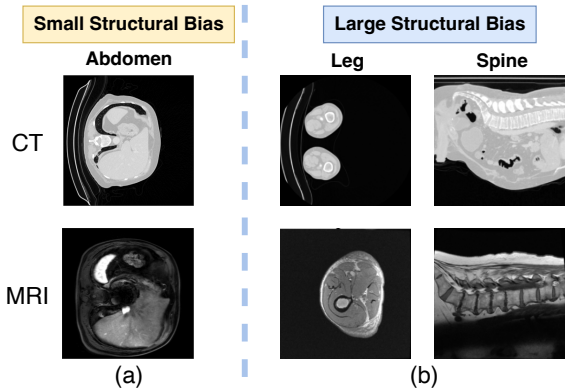

Figure 1: **Structural biases between CT and MRI modalities in certain anatomical regions:** minor for the abdominal region from axial view (a), but severe for the leg from axial view and spinal regions from sagittal view (b).

Ensuring anatomical consistency in unpaired I2I translation is challenging, particularly when the input and output domains exhibit a substantial *structural bias*—i.e., a consistent difference in anatomical structure and shape between domains. An example of this is the drastic visual difference between CT and MRI from different protocols for leg and spinal regions as captured in standard exams (see Fig. 1 and 4), where typically CT images display two legs while MRI scans only show one, and CT images capture entire the abdominal body while MRI focuses on the lumbar area, respectively. Traditional translation models tend to internalize this structural bias, resulting in them applying drastic anatomical transformations during translation in order to align with the typical structure seen in the output domain, resulting in a misalignment between translated images and their corresponding input segmentation masks, potentially leading to unreliable segmentation models trained this data.

One group of methods for unpaired I2I translation in medical imaging is based on Generative Adversarial Networks (GANs) Goodfellow et al. (2020) such as Cycle-consistent Adversarial Network (CycleGAN) Zhu et al. (2017); Armanious et al. (2019); Chen et al. (2023a); Phan et al. (2023); Zhou et al. (2023). These methods maintain the consistency between the images from input and output domains by leveraging cycle consistency loss, minimizing information loss during bidirectional translation Zhu et al. (2017). However, such cycle-consistent supervision does not provide a direct and interpretable constraint on preserving anatomical structures between modalities. Indeed, CycleGAN and its variants may yield undesirable results when substantial misalignment exists between modalities Phan et al. (2023).

Recently, several conditional diffusion models have been introduced for image translation tasks, both in natural images Batzolis et al. (2021); Rombach et al. (2022); Li et al. (2023a); Kim et al. (2023) and medical imaging Li et al. (2023b); Özbey et al. (2023); Kim and Park (2024). However, some of these methods are constrained to paired data or aligning features in domains that are difficult to interpret for unpaired data, such as latent or frequency domains.

To preserve anatomical structures using pixel-level constraints, inspired by previous works in spatially-conditioned diffusion models Konz et al. (2024); Zhang et al. (2023); Rombach et al. (2022), we propose a diffusion model for image translation, "**ContourDiff**" [1], that uses domain-invariant anatomical contour representations of images to guide the translation process, which enforces precise anatomical consistency even between modalities with severe structural biases. This model also has the added benefit of **allowing zero-shot learning**: it solely requires a set of unlabeled output domain images for training, unlike most unpaired translation models. As such, it can potentially translate images from arbitrary unseen domains at inference (see Section 4.7), which can be advantageous for medical image harmonization across multiple imaging modalities. We evaluate our method on CT to MRI translation for sagittal-view lumbar spine and axial-view hip-and-thigh body regions, which both possess severe structural biases (Fig. 1 and 4). In addition to utilizing standard unpaired image generation quality metrics like FID and KID, we evaluate the anatomical consistency of our translation model by training a segmentation model on CT images translated to MRI given their original masks, and evaluating it for real MRI segmentation. Our main contributions include:

1. We propose ContourDiff, a novel diffusion-based framework for unpaired image-to-image translation which allows zero-shot learning.

2. We introduce Spatially Coherent Guided Diffusion (SCGD) to enforces spatial consistency within a volume by providing context information from adjacent slices.

3. Our method significantly outperforms existing unpaired I2I models, including GAN-based and diffusion-based methods, in segmentation performance over all test datasets, despite the fact that it requires no input domain information for training, unlike the competing methods.

4. Our method achieves the best performance compared to existing I2I models in terms of foreground FID and KID across almost all situations.

5. We demonstrate the zero-shot capability of ContourDiff by translating additional input-domain modalities to the output domain without any model retraining.

1. Code: `https://github.com/mazurowski-lab/ContourDiff`

2. Related Works

2.1 Image-to-Image Translation

Image-to-image translation aims to learn a mapping to transform images from one domain to another while preserving essential structural details. Several GAN-based frameworks, including Pix2Pix Isola et al. (2017) and its variants Wang et al. (2018), have been developed as supervised learning methods for paired image-to-image translation. GAN-based models are also widely used in unpaired translation, with CycleGAN Zhu et al. (2017) introducing cycle-consistency loss to allow translation between unpaired datasets. MUNIT Huang et al. (2018) enables multi-modal outputs to generate diverse outputs given images from input domains. GcGAN Fu et al. (2019) incorporates geometric-consistency constraints to preserve the geometric information across domains. To reduce the training time, CUT Park et al. (2020) leverages contrastive learning to align corresponding patches between domains in feature space, instead of using entire images. Despite the success, GAN-based techniques often face challenges like training instabilities and mode collapse problems Li et al. (2023a). More recently, diffusion-based translation frameworks have emerged as a promising alternative, providing competitive performance in both paired Li et al. (2023a) and unpaired Kim et al. (2023) image translation tasks.

Image-to-image translation specialized for medical imaging aims to convert images between modalities (e.g., CT to MRI) to generate synthetic data and improve diagnostic capabilities. However, acquiring labeled and paired medical images is both challenging and expensive Chen et al. (2023b), which exacerbates the challenge of preserving anatomical structures—an essential aspect in medical image translation. To address this issue, several GAN-based frameworks have been developed for *unpaired* medical image translation Armanious et al. (2019); Uzunova et al. (2020); Kong et al. (2021). Recently, diffusion models have gained popularity in this domain. For instance, SynDiff Özbey et al. (2023) incorporates the adversarial diffusion modeling to achieve unsupervised medical image translation. However, these methods rely on adversarial training to align features, lacking strict and interpretable constrains on the detailed anatomical structures during translation.

2.2 Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020), or just *diffusion models*, have recently gained significant attention for their remarkable performance in generative modeling across both natural Croitoru et al. (2023); Müller-Franzes et al. (2023) and medical imaging tasks Rombach et al. (2022); Konz et al. (2024). Different from GAN-based models, diffusion models generate high-quality images with progressive denoising steps, starting from random noise and gradually refining it into a coherent image. Conditional diffusion models extend this approach by incorporating additional conditions, such as texts and images, into the training objectives and model input. For instance, Konz et al. Konz et al. (2024) guided the generation process of medical images with pixel-level masks at each denoising step to ensure strict spatial control over the output. Latent Diffusion Models (LDMs) Rombach et al. (2022) on the other hand shift the diffusion process to a lower-dimensional latent space rather than operating in pixel space for better computational scaling to large images; however, working in this latent space requires a loss of fine detail in the images which the model is conditioned on (in our case, the anatomical contour map) due to downsampling, so our approach remains in image space. Conditional diffusion models have also been explored for other image-to-image tasks, including inpainting Rombach et al. (2022); Corneanu et al. (2024), super-resolution Saharia et al. (2022); Gao et al. (2023) and semantic segmentation Tan et al. (2022); Baranchuk et al. (2021).

3. Methods

3.1 Problem Definition

In unpaired image translation, only unpaired datasets of input and output domain examples are available for training. Our method is even more general in that it accomplishes *zero-shot* image translation, where only an unlabeled dataset of N_{out} output domain examples $[x^{out}]_n$ ($n = 1, \ldots, N_{out}$) are available to train on. The goal is then to use the trained model at inference to translate unseen input domain data $[x^{in}]_n$ to the output domain. In our case, we aim to translate CT images to the MRI domain, for usage with MRI-trained segmentation models. To do so, we propose a novel diffusion-based image translation framework based on domain-invariant anatomical contours of images.

3.2 Adding Contour Guidance to Diffusion Models

3.2.1 Diffusion Models

Denoising diffusion probabilistic models Ho et al. (2020) are generative models that learn to reverse a gradual process of adding noise to an image over many time steps $t = 0, \ldots, T$. New images can be generated by starting with a (Gaussian) noise sample x_T and iteratively applying the model to obtain x_{t-1} from x_t for $t = T, \ldots, 0$ until an image x_0 is recovered.

In practice, the neural network itself $\epsilon_\theta(x_t, t)$ is an I2I architecture (e.g., a UNet Ronneberger et al. (2015)) that is trained to predict the noise ϵ added to an image x_0 at various timesteps t. The training objective is to optimize the Evidence Lower Bound (ELBO). The loss can be simply

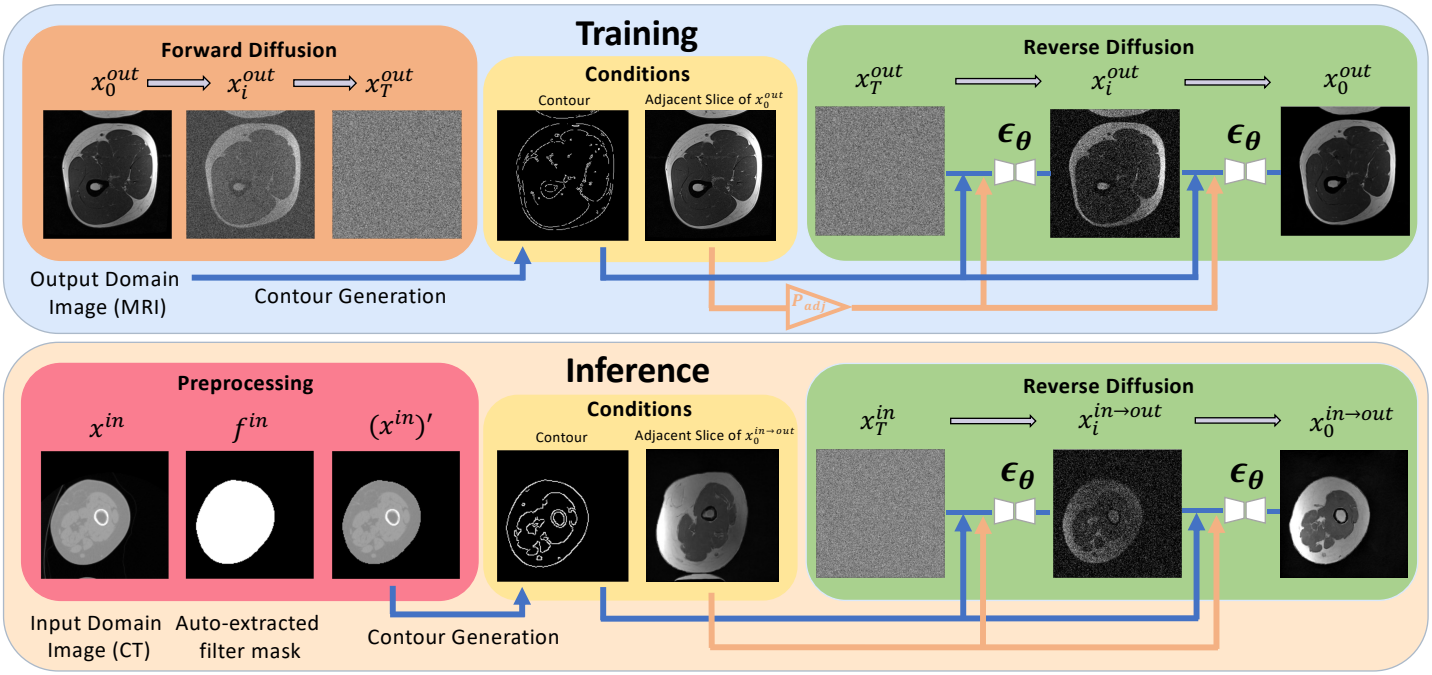

Figure 2: **Overview of ContourDiff.** Top is the training process of ContourDiff. The denoising model ϵ_θ is trained on output domain images, conditioning on their anatomical contours and on an adjacent slice with probability P_{adj}. Bottom is the inference process of ContourDiff. The model generates input domain images in the appearance of the output domain given input domain contours and previously generated adjacent slices.

described as Nichol and Dhariwal (2021):

$$L = \mathbb{E}_{x_0,t,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t)||^2\right] \tag{1}$$

where θ is the model parameters.

Unlike unconditional DDPMs, many conditional diffusion models Rombach et al. (2022); Li et al. (2023a); Konz et al. (2024) directly integrate the conditions y (e.g., images and texts) into the training objective:

$$L = \mathbb{E}_{(x_0,y),t,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t|y)||^2\right], \tag{2}$$

which allows the model to leverage external information to guide the generation process.

Denoising Diffusion Implicit Models (DDIMs) Song et al. (2020) employ a deterministic, non-Markovian sampling process, allowing for faster sample generation without noticeable compromises for image fidelity.

3.2.2 Contour-guided Diffusion Models

For standard unconditional diffusion models, it is unclear how to constrain the semantics/anatomy of generated images. To address this, we propose to utilize *contour* representations of images to provide guidance in generating the image. While training the model, we use the Canny edge detection filter Canny (1986) to extract the contour representation c of each training image x_0, similar as that in Rombach et al. (2022), and concatenate it with the network input at every denoising step, a practice similar to Konz et al. (2024); Zhang et al. (2023). This modifies the network in Eq. 2 to become $\epsilon_\theta(x_t, t|c)$ and the diffusion training objective to become

$$L = \mathbb{E}_{(x_0,c),t,\epsilon}\left[||\epsilon - \epsilon_\theta(x_t, t|c)||^2\right], \tag{3}$$

where (x_0, c) is a training set image and its accompanying contour. We perform this in image space in order to ensure that the denoised image precisely follows the contour guidance pixel-to-pixel (as in Konz et al. (2024)), which may be lost if diffusion is performed within a latent space Rombach et al. (2022).

3.3 Contour-guided image translation

3.3.1 Overall Translation Process

One important feature of contours is that they can be viewed as domain-invariant yet anatomy-preserving representations of images. This allows for a contour-guided diffusion model trained in some output domain to serve as a zero-shot image translation method, as follows.

First, we train a contour-guided diffusion model on output domain images with accompanying contours $([x^{out}]_n, [c^{out}]_n)$, shown in Algorithm 1 (Note: Algorithm 1 also include constraints from adjacent slices). Next, to translate some *input*

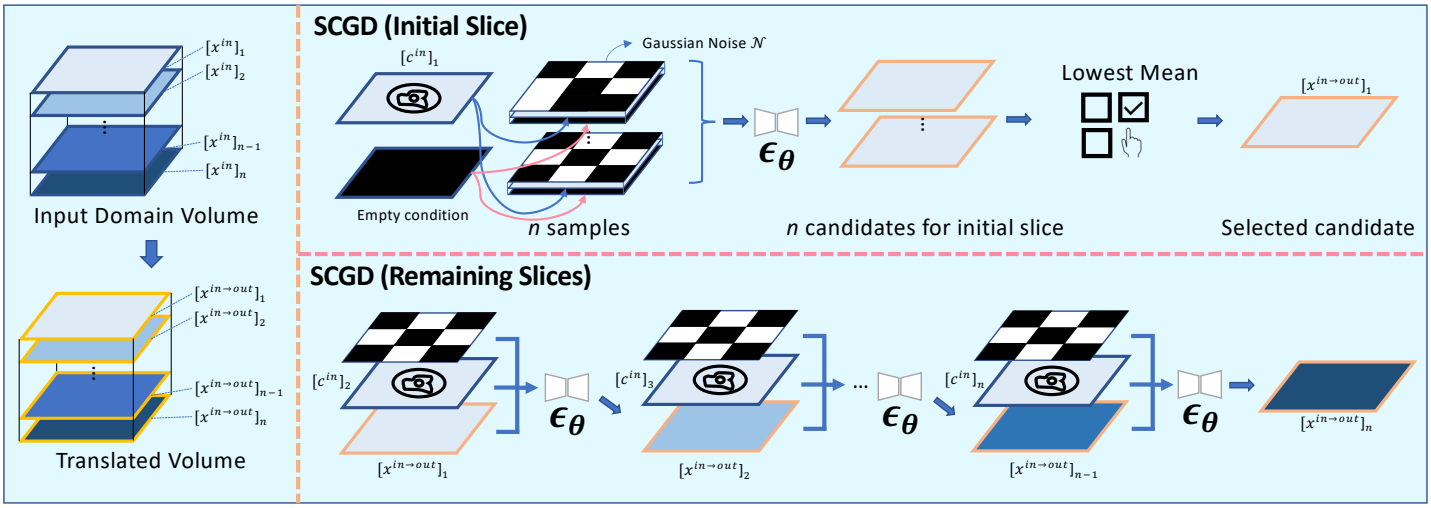

Figure 3: **Spatially Coherent Guided Diffusion (SCGD).** For each input domain volume, SCGD first translates the initial slice by generating n candidates with setting C_{adj} to an empty map and selecting the optimal one according to a specified criterion (e.g., lowest mean intensity). Then, every subsequent slice is synthesied by conditioning on its anatomical contours and the previously translated slice. Input domain slices are bordered in blue and output domain slices are bordered in orange.

domain image x^{in} to the output domain, we extract its contour c^{in} after removing irrelevant backgrounds using F_{filter}, and use the output domain-trained model ϵ_θ conditioned on c^{in} to generate the image $x^{in\to out}$. Therefore, $x^{in\to out}$ maintains the anatomical content of x^{in}, while possessing the visual domain characteristics of the output domain. Our translation algorithm is shown in Algorithm 2, where $\alpha_t = 1 - \beta_t$ with the variance of the additive pre-scheduled noise β_t, and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

3.3.2 Filtering Out Image Artifacts

We also apply additional pre-processing to network input images x to filter out non-anatomical features/artifacts (e.g., the motorized table in CT) if necessary, by applying a binary mask M_{filter} as $x \leftarrow M_{filter} \odot x$. M_{filter} is defined by sequentially computing the follow Scikit-Image van der Walt et al. (2014) functions on x Phan et al. (2023): `threshold_multiotsu`, `binary_erosion`, `remove_small_objects`, and `remove_small_holes`.

3.4 Spatially Coherent Guided Diffusion (SCGD)

We introduce Spatially Coherent Guided Diffusion (***SCGD***), a novel framework designed to preserve spatial consistency when translating adjacent slices from 3D volumes (e.g., CT) into an output domain. SCGD jointly leverages anatomical contour information from the current slice and spatial contextual guidance from its neighbors to enforce both spatially coherent and anatomically consistent translations for each volume. To enable such joint conditioning, SCGD introduces an additional input channel for the diffusion model, C_{adj}.

3.4.1 Training

During training, each reverse diffusion step is conditioned on the contour of the current slice, $[x^{in}]_i$, together with one adjacent slice. As slice ordering may be indeterminate at inference, we randomly choose either $[x^{in}]_{i+1}$ or $[x^{in}]_{i-1}$ with equal probability to enable bidirectional spatial guidance.

Adjacent Slice Ratio in Training To ensure the model learns contour-based image generation, we incorporate adjacent slices as conditioning inputs with probability P_{adj}. For each training datapoint, C_{adj} is set as:

$$C_{adj} = \begin{cases} [x_0^{out}]_{adj}\text{ , with probability } P_{adj} \\ \mathbf{0}_{2D}\text{ , with probability } 1-P_{adj} \end{cases} \tag{4}$$

where $[x_0^{out}]_{adj}$ is the adjacent output-domain slice and $\mathbf{0}_{2D}$ is the empty condition map. Intuitively, the model should rely more on anatomical contour information from the current input slice than on information from adjacent slices. To enforce this, we set P_{adj} no larger than 0.5. This choice is particularly important for translating the initial slice for each volume at inference, where no adjacent translated slice is available. Moreover, keeping P_{adj} moderate can prevent the accumulation of errors across slices.

Algorithm 1 ContourDiff Training Phase

Input: Output domain training distribution $p(x_0^{out})$.
repeat
  $x_0^{out} \sim p(x_0^{out})$
  $c^{out} = \texttt{Canny}(x_0^{out})$
  if *rand()* $\leq P_{adj}$ **then**
    $[x_0^{out}]_{adj} = \texttt{Slice adjacent to } x_0^{out}$
  else
    $[x_0^{out}]_{adj} = \mathbf{0}$
  end
  $\epsilon \sim \mathcal{N}(0, I_n)$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $x_t^{out} = \sqrt{\bar{\alpha}_t} x_0^{out} + \sqrt{1-\bar{\alpha}_t}\epsilon$
  Update θ with $\nabla_\theta \left\| \epsilon - \epsilon_\theta(x_t^{out}, t | (c^{out}, [x_0^{out}]_{adj}) \right\|^2$
until *converged*;

Algorithm 2 ContourDiff Inference Phase

Input: Input domain image x^{in}.
Output: Translated image $x_0^{in \to out}$
$c^{in} = \texttt{Canny}(x^{in})$
$x_T^{out} \sim \mathcal{N}(0, I_n)$
if *Initial Slice* **then**
  $[x_0^{in \to out}]_{adj} = \mathbf{0}$
else
  $[x_0^{in \to out}]_{adj} = \texttt{Slice adjacent to } x_0^{in \to out}$
end
for $t = T, \ldots, 1$ **do**
  $\epsilon \sim \mathcal{N}(0, I_n)$ if $t > 1$, else $\epsilon = 0$
  $x_{t-1}^{out} = \frac{1}{\sqrt{\alpha_t}}$
  $\times \left(x_t^{out} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t^{out}, t | (c^{in}, [x_0^{in \to out}]_{adj})) \right)$
  $+ \sigma_t \epsilon$
end
return $x_0^{in \to out}$

3.4.2 Inference

As illustrated in the Fig. 3, we first translate the initial slice, $[x^{in}]_1$, of a given 3D volume to its output domain version $[x^{in \to out}]_1$. To obtain a robust starting point, we generate n samples in parallel (we use $n = 16$ in our experiments) and select the sample exhibiting the lowest mean intensity, as we empirically observe that instability in stochastic reverse diffusion can produce overly bright backgrounds. Then, the remaining slices within the volume are translated sequentially, conditioning each step on both the contour of the current input slice and on the previous translated slice to preserve anatomical consistency and spatial coherence throughout the volume. To accelerate the translation, we provide a volume-group parallel inference implementation that processes multiple groups in parallel. Specifically, we partition all volumes evenly into groups and process each group concurrently via scripts. The overall pseudocode for the training and inference phases with SCGD is presented in Algorithms 1 and 2, respectively.

4. Experiment

4.1 Datasets

In this paper, we study one of the most common translation scenarios, CT to MRI, based on three datasets: **TotalSegmentator** public dataset Wasserthal et al. (2023), **SPIDER** lumbar spine (L-SPIDER) public dataset van der Graaf et al. (2023) and a private in-house dataset.

For the MRIs used to train the ContourDiff, we collect a private dataset with T1 weighted lumbar spine (L) and hip & thigh (H&T) body regions. 40 sagittal lumbar MRI volumes (670 2D slices), and 10 axial MRI volumes from thigh and hip (404 2D slices) are selected. Correspondingly, we obtain 54 sagittal (2,333 2D slices) and 29 axial (4,937 2D slices) CT volumes from the TotalSegmentator Wasserthal et al. (2023) in L and H&T, respectively.

For downstream bone segmentation task, we further randomly split the two CT sets by patients (43:11 for L and 23:6 for H&T) for training and validation. We evaluate the segmentation performance on held-out annotated MRI sets (10 L volumes including 158 2D slices, 12 H&T volumes including 426 2D slices). In addition, to study the generalization ability of our method, we test the lumbar segmentation model on 40 volumes (731 2D slices) from L-SPIDER van der Graaf et al. (2023) [2]. Moreover, we collect an additional held-out annotated MRI dataset to train the segmentation model directly on real output-domain images (352:80 2D slices for L; 990:305 2D slices for H&T), which serves as an upper bound (UB) for performance.

4.2 Evaluation Metrics

We quantitatively evaluate translation performance by first training segmentation models on translated images with input domain (CT) masks and testing on real output domain (MRI) images. We adopt commonly-used metrics, Dice Coefficient (DSC) and average symmetric surface distance (ASSD), both evaluated on 3D volumetric segmentation. Given the predicted mask A and the ground truth mask B, the two metrics are defined as:

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{5}$$

DSC measures the overlap between A and B, ranging from 0 and 1. Higher DSC represents better segmentation

2. We crop the slices to exclude the sacrum, as it is not annotated.

performance.

$$ASSD(A,B) = \frac{\sum_{p\in S_A} d(p,S_B) + \sum_{q\in S_B} d(q,S_A)}{|S_A| + |S_B|} \quad (6)$$

where S_A and S_B are the sets of surface points of mask A and B. $d(p,S_B)$ and $d(q,S_A)$ are the shortest Euclidean distances from a surface point in S_A and S_B to the nearest point in S_B and S_A, respectively. ASSD penalizes more on boundary errors relative to DSC. ASSD is non-negative and lower ASSD values indicate better segmentation performance.

In addition, we evaluate the boundary alignment between the Canny edges extracted from the input-domain 2D images and the translated 2D images using the 95th percentile Hausdorff Distance (HD95). Given edge sets A (input-domain) and B (translated), HD95 is defined as:

$$HD95(A,B) = \max\Big(\mathrm{P}_{95}(d(A,B)), \mathrm{P}_{95}(d(B,A))\Big) \quad (7)$$

where $d(A,B) = \{\min_{b\in B} \|a-b\| \mid a \in A\}$ denotes the set of distances from each point in A to the nearest point in B. HD95 is non-negative, with smaller values indicating better contour alignment.

As there are no paired images, we also calculate the foreground[3] FID Heusel et al. (2017) and KID Bińkowski et al. (2018) between the translated image and output domain image distributions for reference. We do this to measure the feature alignment of the foreground object between input and output domains, free of noise from the surrounding background areas which are less important for the segmentation tasks of interest.

4.3 Implementation Details

4.3.1 Model Architecture

For the image translation model, we adopt the UNet architecture Ronneberger et al. (2015) for the denoising model ϵ_θ with a three-channel input (grayscale image, its contour and spatial information from a grayscale neighbor slice). The encoder comprises six down sampling stages, each consisting of two ResNet blocks. The number of output feature channels at each stage is (128, 128, 256, 256, 512, 512). To capture long-range dependencies, we integrate spatial self-attention at the fifth stage. The decoder symmetrically upsamples through six stages, fusing encoder features via skip connections, and produces a single-channel translated image at the same resolution as the input. For canny edge extraction, we inspected a representative subset of images and selected low/high thresholds of 30/50 for CT and 50/100 for MRI to optimally capture anatomical contours while suppressing spurious edges. We primarily conduct experiments with $P_{adj} = 0.2$ and evaluate other values in the ablation studies. All experiments are running on a NVIDIA RTX A6000 GPU.

4.3.2 Model Training

The training settings for the diffusion model follow the same as that in Konz et al. (2024). We use the DDIM algorithm Song et al. (2021) for sampling, with 50 steps. For the segmentation models, we use the convolution-based UNet Ronneberger et al. (2015) and transformer-based SwinUNet Cao et al. (2022). All images are resized to 256×256 and normalized to 8-bit $[0, 255]$, following common preprocessing practices in medical imaging analysis Mazurowski et al. (2023); Ma et al. (2024); Lyu et al. (2024); Konz et al. (2024). The training of competing methods mostly follows the default settings from each official GitHub [4]. We set the $\lambda_{idt} = 0.5$ to include identity loss if provided. We train the segmentation model with a cosine learning rate scheduler up to 100 epochs with the initial learning rate of 1×10^{-3}.

4.4 Comparison with Other Methods

We compare our framework to other translation/adaptation methods, including CycleGAN Zhu et al. (2017), SynSeg-Net Huo et al. (2019), CyCADA Hoffman et al. (2017), MUNIT Huang et al. (2018), CUT Park et al. (2020), GcGAN Fu et al. (2019), MaskGAN Phan et al. (2023), FGDM Li et al. (2023b) and UNSB Kim et al. (2023), via the performance of output domain-trained downstream task segmentation models on translated images. Several of these methods (e.g., Zhu et al. (2017); Huo et al. (2019); Huang et al. (2018); Park et al. (2020); Fu et al. (2019); Phan et al. (2023); Li et al. (2023b); Kim et al. (2023)) translate the images solely at the image level, while CyCADA also aligns latent features from the downstream task encoder. MaskGAN uses the extracted coarse masks to better preserve object structures throughout translation. In addition to GAN-based models, FGDM utilizes both low and high frequency information as diffusion conditions for translation, and UNSB integrates diffusion models with Schrödinger Bridge theory to enable probabilistically consistent translation for unpaired data. For CyCADA, we used the same segmentation architecture as other methods but without the skip connection to enable feature alignment. For each competing method, we evaluated multiple intermediate results from the translation [5] and report the best performance.

3. Foreground refers to pixels containing the object of interest. In this paper, we use masks from CTs to extract objects.

4. Training time: 200 epochs (CycleGAN, SynSeg-Net, CyCADA, GcGAN, MaskGAN); 400 epochs (CUT); 1M iterations (MUNIT); 60/280 epochs (UNSB on L/H & T).

5. For SynSeg-Net and CyCADA, we evaluate the segmentation model every 20 epochs. For other methods, as we need to train the segmentation model separately, we evaluate at 10%, 30%, 50%, 75% and 100% of the training time.

	Lumbar					Lumbar-SPIDER					Hip & Thigh				
	UNet		SwinUNet			UNet		SwinUNet			UNet		SwinUNet		
Method	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	Edge HD95 (↓)	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	Edge HD95 (↓)	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	Edge HD95 (↓)
w/o Adap.	0.287 ± 0.034	6.515 ± 1.495	0.171 ± 0.039	7.386 ± 1.263	-	0.236 ± 0.023	8.275 ± 0.654	0.187 ± 0.022	8.327 ± 0.600	-	0.004 ± 0.002	45.731 ± 5.724	0.003 ± 0.002	48.624 ± 5.429	-
CycleGAN	0.484 ± 0.022	2.479 ± 0.160	0.362 ± 0.028	3.505 ± 0.259	24.673 ± 0.189	0.507 ± 0.015	3.629 ± 0.284	0.412 ± 0.021	3.701 ± 0.263	24.673 ± 0.189	0.535 ± 0.038	9.140 ± 1.642	0.464 ± 0.053	9.790 ± 1.247	13.541 ± 0.117
SynSeg-Net	0.316 ± 0.031	3.013 ± 0.410	0.288 ± 0.040	3.527 ± 0.358	26.737 ± 0.293	0.364 ± 0.019	3.207 ± 0.197	0.223 ± 0.023	7.366 ± 0.659	26.737 ± 0.293	0.370 ± 0.064	4.705 ± 0.456	0.059 ± 0.014	12.871 ± 1.469	20.123 ± 0.262
CyCADA	0.331 ± 0.024	5.942 ± 1.219	0.319 ± 0.024	3.691 ± 0.260	30.109 ± 0.262	0.364 ± 0.016	4.389 ± 0.256	0.260 ± 0.011	4.725 ± 0.186	30.109 ± 0.262	0.349 ± 0.039	11.247 ± 1.472	0.155 ± 0.033	13.002 ± 1.684	12.544 ± 0.105
MUNIT	0.407 ± 0.013	3.803 ± 0.223	0.433 ± 0.016	3.212 ± 0.213	44.285 ± 0.648	0.380 ± 0.013	4.309 ± 0.290	0.358 ± 0.014	3.545 ± 0.174	44.285 ± 0.648	0.128 ± 0.026	16.228 ± 3.226	0.090 ± 0.023	18.925 ± 3.179	22.358 ± 0.205
CUT	0.392 ± 0.020	4.670 ± 0.745	0.288 ± 0.029	5.259 ± 0.371	32.665 ± 0.702	0.368 ± 0.020	5.781 ± 0.427	0.292 ± 0.022	6.751 ± 0.530	32.665 ± 0.702	0.311 ± 0.052	19.254 ± 3.817	0.211 ± 0.030	20.564 ± 4.384	27.118 ± 0.261
GcGAN	0.554 ± 0.020	1.753 ± 0.087	0.433 ± 0.030	2.940 ± 0.372	11.683 ± 0.216	0.580 ± 0.010	2.202 ± 0.096	0.513 ± 0.013	2.904 ± 0.157	11.683 ± 0.216	0.414 ± 0.048	9.275 ± 2.035	0.320 ± 0.043	13.650 ± 2.459	25.998 ± 0.271
MaskGAN	0.428 ± 0.026	3.192 ± 0.251	0.322 ± 0.039	4.692 ± 0.917	16.961 ± 0.213	0.458 ± 0.017	3.729 ± 0.253	0.385 ± 0.023	5.355 ± 0.438	16.961 ± 0.213	0.289 ± 0.048	16.229 ± 3.576	0.292 ± 0.032	17.590 ± 3.245	30.838 ± 0.276
FGDM	0.455 ± 0.022	4.658 ± 0.727	0.390 ± 0.022	5.589 ± 0.783	15.701 ± 0.167	0.411 ± 0.021	5.077 ± 0.441	0.333 ± 0.025	6.348 ± 0.510	15.701 ± 0.167	0.074 ± 0.019	31.927 ± 5.861	0.070 ± 0.020	29.386 ± 5.170	38.282 ± 0.258
UNSB	0.465 ± 0.028	3.111 ± 0.263	0.456 ± 0.016	2.954 ± 0.206	29.499 ± 0.645	0.488 ± 0.014	3.984 ± 0.437	0.446 ± 0.017	3.070 ± 0.132	29.499 ± 0.645	0.247 ± 0.035	13.426 ± 2.399	0.181 ± 0.041	17.650 ± 3.717	19.101 ± 0.187
Ours	**0.705 ± 0.019**	**1.677 ± 0.507**	**0.669 ± 0.019**	**1.570 ± 0.217**	**3.396 ± 0.020**	**0.655 ± 0.011**	**1.955 ± 0.153**	**0.603 ± 0.016**	**2.226 ± 0.191**	**3.396 ± 0.020**	**0.769 ± 0.036**	**2.625 ± 0.533**	**0.684 ± 0.060**	**4.258 ± 0.977**	**3.578 ± 0.078**
UB†	0.748 ± 0.021	1.296 ± 0.182	0.740 ± 0.021	1.315 ± 0.162	-	0.764 ± 0.005	1.325 ± 0.040	0.765 ± 0.004	1.421 ± 0.057	-	0.857 ± 0.020	1.678 ± 0.283	0.786 ± 0.042	2.992 ± 0.569	-

Table 1: Quantitative comparison (DSC, ASSD and Edge HD95) of ContourDiff to other image translation methods in terms of segmentation model performance on held-out output domain images. (L: Lumbar dataset, L-SPIDER: SPIDER Lumbar dataset, H & T: Hip & Thigh dataset). "w/o Adap." is the baseline referring to the model trained on CTs without any adaptation and tested on MRIs directly. UB† represents the upper bound model trained on real annotated output domain images. Best in bold, runner-up underlined. Standard error of the mean (SEM) is reported with each result.

Lumbar Spine (L) - Foreground										
Metric	CycleGAN	SynSeg-Net	CyCADA	MUNIT	CUT	GcGAN	MaskGAN	FGDM	UNSB	**Ours**
FID (↓)	132.16	137.63	127.54	372.67	150.10	138.60	128.17	158.69	137.42	**126.35**
KID (↓)	0.047	0.054	0.045	0.343	0.058	0.050	**0.039**	0.071	0.051	0.042
Hip & Thigh (H & T) - Foreground										
Metric	CycleGAN	SynSeg-Net	CyCADA	MUNIT	CUT	GcGAN	MaskGAN	FGDM	UNSB	**Ours**
FID (↓)	183.18	192.32	184.11	193.12	193.63	163.61	175.28	251.85	167.88	**133.74**
KID (↓)	0.163	0.169	0.159	0.174	0.178	0.144	0.152	0.257	0.142	**0.093**

Table 2: Quantitative comparison of foreground FID and KID between translated images and output domain images. Best in bold, runner-up underlined. (L-SPIDER is excluded as it is only used for testing and not for training.)

4.5 Results

4.5.1 Quantitative Results

The segmentation model results are shown in Table 1. Overall, across all three test sets, our method consistently outperforms previous image translation methods by a significant margin on both DSC and ASSD, using either a UNet or a SwinUNet. Specifically, as for UNet, our method achieves DSC improvements of at least 15.1%, 7.5% and 23.4% on the L, L-SPIDER and H&T datasets, respectively. When using SwinUNet, the DSC improvements of segmentation model are at least 21.3%, 9% and 22% on the same three datasets. These results demonstrate the superior translation performance of our method in developing cross-domain segmentation models. Furthermore, our method significantly outperforms all baselines in terms of edge alignment, reducing HD95 by 8.287 and 8.966 in compared with the runner-up methods in lumbar and hip & thigh regions, respectively, indicating better anatomical contour alignment during translation.

Based on Table 2, our method achieves the lowest FID scores: 126.35 and 133.74 for L and H & T, respectively. For KID scores, our method outperforms others for H & T and achieves a close second place for L (0.042), which is slightly lower than the top score of 0.039 by MaskGAN. The improvements in FID and KID demonstrate the superior foreground fidelity of the images translated by our method compared to other baselines.

4.5.2 Qualitative Results

We provide example translated images in Fig. 4. These datasets form a challenging task due to (1) the noticeable shift in image features between the input and output domains and (2) the high anatomical variability between different scans. Moreover, we see that adversarially-trained models (e.g., CycleGAN) have trouble with the consistent structural shift (i.e., large structural bias) between the input and output domains, i.e., when one domain is absent of certain features seen in the other. As shown in Fig. 1, this is particularly evident in our H&T dataset, where MRIs are dominant by a single leg, and CTs often contain two legs. Such a bias may lead the adversarial mechanism to overemphasize these features and, therefore, tend to translate CTs of two legs into MRIs depicting only one leg (see Fig. 4). For the lumbar spine from the sagittal view, MRIs often start from the lowest thoracic spine and end at the sacrum.

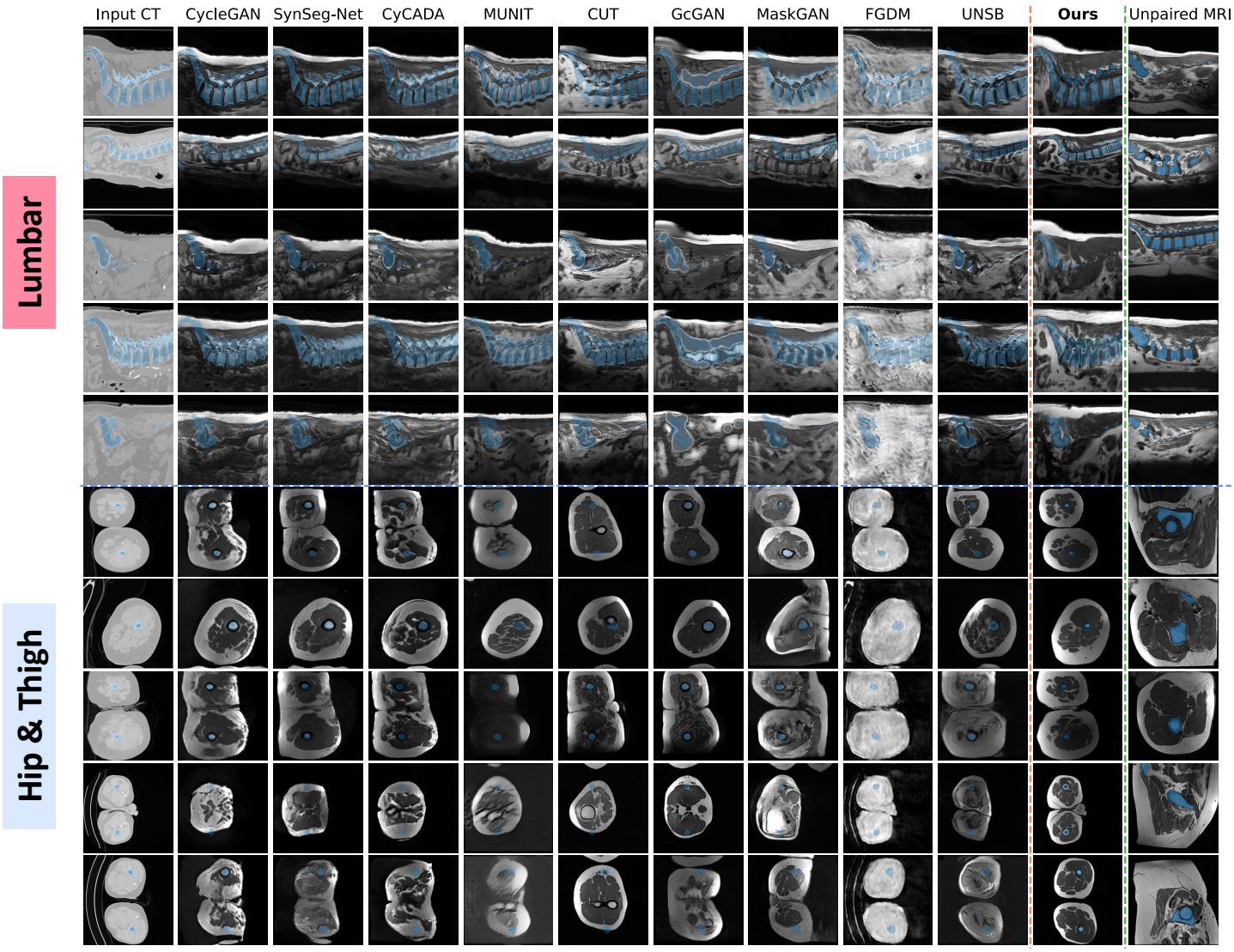

Figure 4: **Qualitative comparison of ContourDiff and baseline methods.** ContourDiff appears to best maintain anatomical consistency during translation for both Lumbar and Hip & Thigh areas. The input-domain segmentation masks are depicted in blue to visualize the alignment. Unpaired MRIs are included as target-domain examples for reference only. Note: they are no used as ground truth for the translation.

On the other hand, CTs often include the upper leg and sometimes the abdominal body (see Fig. 4).

Fig. 4 shows that our model explicitly enforces anatomical consistency through translation despite these domain feature differences through its contour guidance, generating MRIs that strictly follow input CT images, resulting in better mask alignment and better segmentation model performance. Notably, the translated outputs from ContourDiff also preseve clinically relevant anatomical details, such as clear boundaries between fat and muscle, as well as other major structures.

Based on Table 1, Table 2 and Fig. 4, ContourDiff best maintain anatomical fidelity and consistency compared to other models, both quantitatively and qualitatively.

4.6 Ablation Studies

We will now conduct ablation studies to validate the effectiveness of key design choices for ContourDiff, studying how contours are used for guidance, and the general effectiveness of SCGD and its dependence on P_{adj}.

4.6.1 Effectiveness of Adding Contours

We verify the effectiveness of introducing contours to each denoising step during training by conditionally training on an empty map (i.e., all zeros) and adding the CT contours during the translation steps. Fig. 5 showed that the denoised model ϵ_θ trained without contours hardly followed the introduced CTs contours (**'Uncond.'** column). Furthermore, the UNet trained on these unconditionally

generated MRIs experienced a dramatic performance drop (see Table 3). These results demonstrate the necessity of including contours to achieve anatomical consistency during translation.

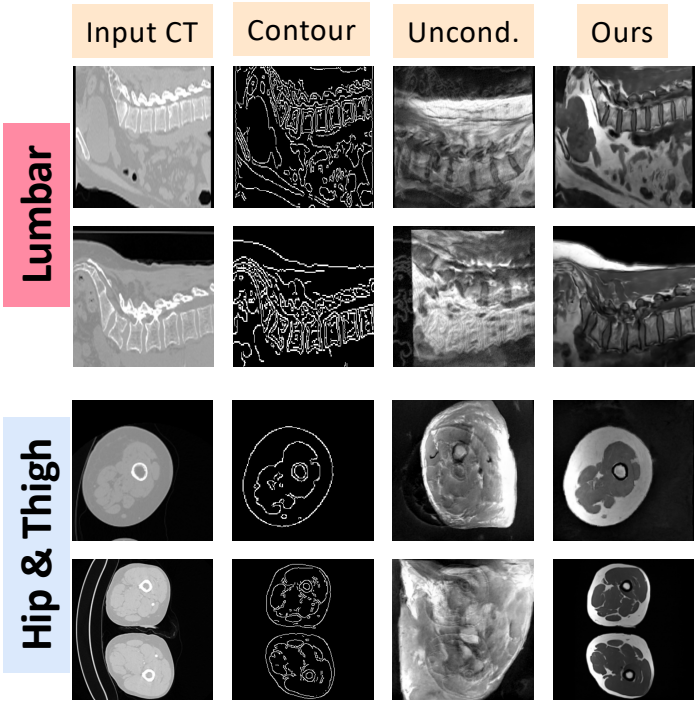

Figure 5: **Qualitative comparison between unconditional DDPM and ContourDiff.** Unconditional DDPM seems to hardly follow input-domain anatomical structures during translation.

	L		L-SPIDER		H & T	
Method	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)
Unconditional	0.264 ± 0.044	5.148 ± 1.337	0.202 ± 0.023	6.479 ± 0.595	0.298 ± 0.032	19.352 ± 4.153
Ours	**0.705 ± 0.019**	**1.677 ± 0.507**	**0.655 ± 0.011**	**1.955 ± 0.153**	**0.769 ± 0.036**	**2.625 ± 0.533**

Table 3: Quantitative comparison (DSC and ASSD) of ContourDiff and unconditional DDPM. Best in bold.

4.6.2 Effectiveness of SCGD

To assess the impact of the proposed SCGD, we also generate translated images without guidance from the adjacent slice. Similarly, we then train a UNet and report the segmentation performance with the standard error of the mean (SEM). As shown in Table 4 and Fig. 6, ContourDiff with SCGD achieves a superior performance compared to that without SCGD (i.e., without guidance from adjacent slices). Moreover, according to Table 1 and 4, ContourDiff without SCGD still outperforms baseline methods by a significant margin, further demonstrating its promise for preserving anatomical structures with 2D data.

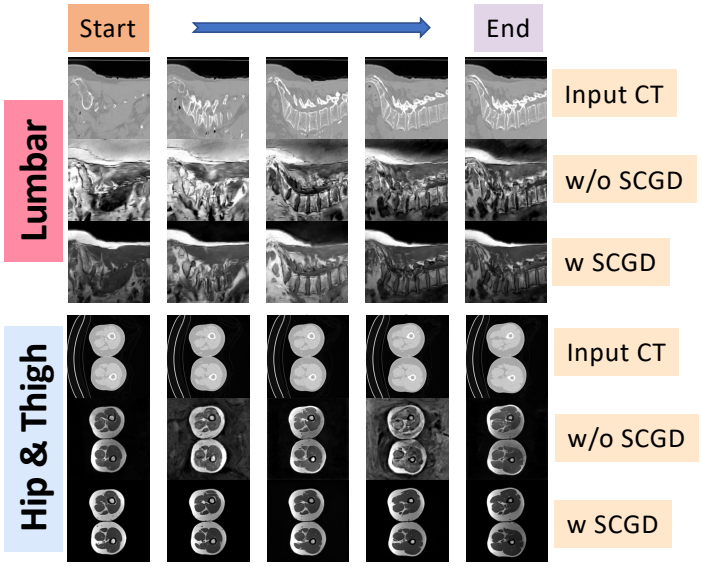

Figure 6: **Qualitative results of translation with SCGD.** SCGD can better preserve the consistency between translated slices within volume.

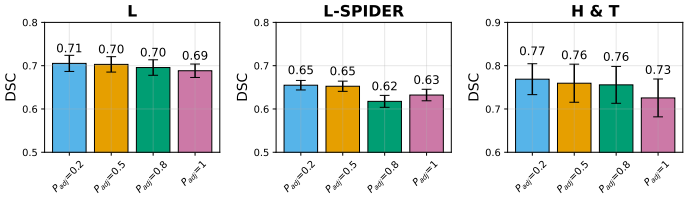

Figure 7: **Quantitative comparison of different choices on P_{adj}.** Smaller P_{adj} seems to result in better translation performance.

	L		L-SPIDER		H & T	
Method	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)
w/o SCGD	0.653 ± 0.025	2.425 ± 0.758	0.583 ± 0.021	2.315 ± 0.221	0.706 ± 0.044	3.489 ± 0.540
Ours	**0.705 ± 0.019**	**1.677 ± 0.507**	**0.655 ± 0.011**	**1.955 ± 0.153**	**0.769 ± 0.036**	**2.625 ± 0.533**

Table 4: Quantitative comparison (DSC and ASSD) of ContourDiff with SCGD and without SCGD. Best in bold.

4.6.3 Different Choices on P_{adj}

We now investigate how varying P_{adj} affects translation quality by training and evaluating the UNet architecture for each setting. As shown in Table 5 and Fig. 7, lower P_{adj} values seem to result in better translation performance. This outcome is expected because a smaller P_{adj} forces the model to rely more heavily on anatomical contour constraints. By contrast, a larger P_{adj} may introduce conflicting contextual information from adjacent slices and may encourage the network to learn direct mappings from adjacent slices instead of preserving true anatomical fidelity from contours. Thus, in practice, we recommend choosing P_{adj} at or below 0.5.

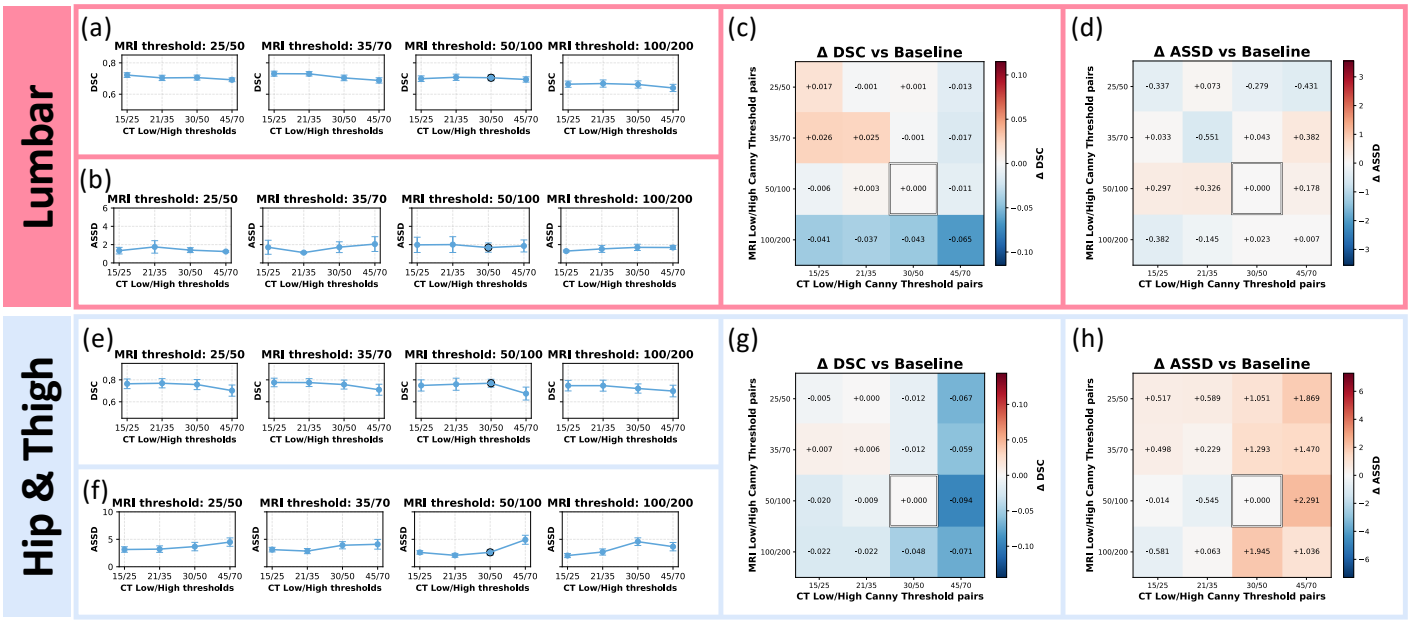

Figure 8: **Model performance under different Canny threshold pairs.** (a), (e) and (b), (f) show DSC and ASSD with standard error of the mean (SEM) across threshold pairs. (c), (g) and (d), (h) present heatmaps of the corresponding DSC and ASSD deltas relative to the baseline setting.

	L		L-SPIDER		H & T	
P_{adj}	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)	DSC (↑)	ASSD (↓)
1	0.688 ± 0.016	1.555 ± 0.284	0.632 ± 0.013	2.279 ± 0.203	0.726 ± 0.044	4.480 ± 0.732
0.8	0.696 ± 0.018	**1.194 ± 0.099**	0.618 ± 0.014	**1.902 ± 0.127**	0.756 ± 0.043	3.285 ± 0.637
0.5	0.703 ± 0.018	1.393 ± 0.277	0.653 ± 0.012	1.949 ± 0.147	0.760 ± 0.044	2.639 ± 0.670
Ours	**0.705 ± 0.019**	1.677 ± 0.507	**0.655 ± 0.011**	1.955 ± 0.153	**0.769 ± 0.036**	**2.625 ± 0.533**

Table 5: Quantitative results (DSC and ASSD) of ablation study in terms of segmentation model performance to explore different P_{adj} values. Best in bold, runner-up underlined.

	H & T (T2 MRI→T1 MRI)		
Method	DSC (↑)	ASSD (↓)	Edge HD95 (↓)
w/o Adap.	0.647 ± 0.062	7.810 ± 2.313	-
CycleGAN	0.714 ± 0.047	4.669 ± 0.811	25.418 ± 0.708
UNSB	0.724 ± 0.048	4.366 ± 0.938	18.649 ± 0.706
Ours	**0.778 ± 0.040**	**2.927 ± 0.472**	**5.699 ± 0.154**
UB†	0.857 ± 0.020	1.678 ± 0.283	-

Table 6: Quantitative comparison (DSC, ASSD and Edge HD95) of ContourDiff on translating T2-weighted MRI to T1-weighted MRI. Best in bold, runner-up underlined.

4.7 Experiments on T2 MRI to T1 MRI

To further verify the zero-shot capability of ContourDiff, we incorporate an additional experiment on translating T2-weighted MRI to T1-weighted MRI in the hip and thigh region. We collect 594 annotated 2D T2 MRI slices, and split into 502:92 for training and validation. We directly apply the ContourDiff model previously trained for CT to MRI translation to perform T2 to T1 translation in a zero-shot manner, followed by downstream segmentation evaluation using UNet on the same test set as used for CT to MRI tasks.

For comparison, we include representative GAN-based (CycleGAN) and diffusion-based (UNSB) baselines. As shown in Table 6, ContourDiff model not only achieves superior segmentation performance on T2 to T1 translation but also best aligns edges, all without any additional training, demonstrating its zero-shot capability.

4.8 Model Robustness Evaluation

We evaluate the robustness of ContourDiff in terms of different Canny thresholds, image qualities, and image contrasts.

4.8.1 Model Robustness under Different Canny Thresholds

We select three threshold pairs around the current settings for both CTs (30/50) and MRIs (50/100), and evaluate segmentation performance using UNet on both L and H&T datasets. Specifically, we consider low/high threshold pairs of 15/25, 21/35, and 45/70 for CT, and 25/50, 35/70, and 100/200 for MRI.

As shown in Figure 8, ContourDiff mostly exhibits robust performance across different Canny threshold pairs. Lower thresholds, which produce more detailed contour maps, seem to generally maintain or even slightly improve performance. In contrast, higher thresholds reduce the number of detected contours and occasionally lead to small performance drops. This is reasonable as sparse edge information may weaken anatomical guidance. In particular,

even with minor performance degradation, ContourDiff still significantly outperforms all existing baselines (compared to Table 1).

4.8.2 Model Robustness under Different Image Qualities

We simulate variations in image quality by adding Gaussian noise to the original images at signal-to-noise ratio (SNR) levels. Specifically, given a desired SNR in decibels (SNR_{dB}), the noise power is:

$$P_{noise} = \frac{P_{signal}}{10^{\frac{SNR_{dB}}{10}}} \tag{8}$$

where P_{signal} is the mean sqaured intensity of the original image and P_{noise} is the variance of added noise. We add noise at three SNR levels (30 dB, 25 dB, and 15 dB) to the input CT images and evaluate segmentation performance using UNet for both L and H&T datasets.

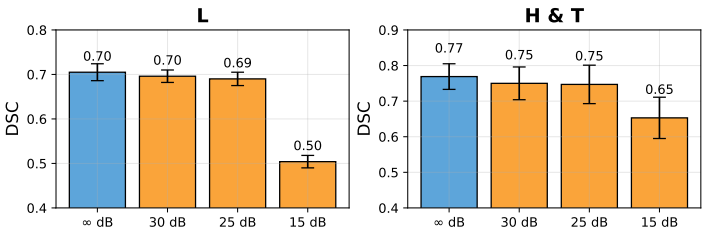

Figure 9: **Model performance under different image qualities.** ∞ dB represents the original images without any added noise.

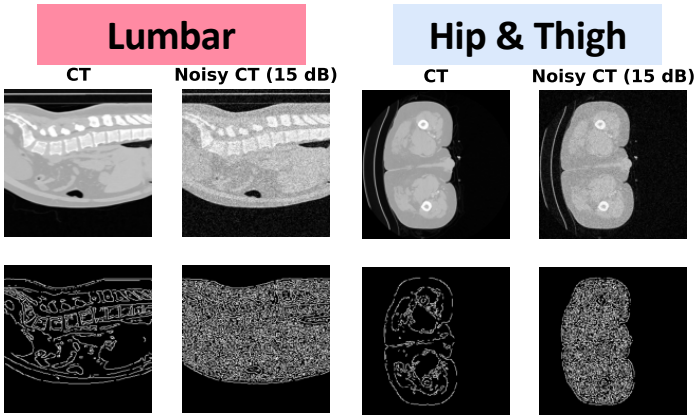

Figure 10: **Image examples and extracted contours under severe noise.** Contours extracted from heavily degraded images tend to lose anatomical details.

Based on Figure 9, the ContourDiff model remains robust under mild-to-moderate noise levels (e.g., 30 dB and 25 dB), with at most a 0.02 DSC drop. Moreover, the performance degrades under more severe noise (e.g., 15 dB), as the extracted contours under such conditions contain limited anatomical details (Figure 10).

4.8.3 Model Robustness under Different Image Contrasts

We adjust contrasts of the CT images and MRI images to assess model robustness. A linear contrast transformation is applied to the original images as follows:

$$I^{'} = c + k \times (I - c) \tag{9}$$

where I is the original pixel intensity, c is the mean intensity of each image, and k is the contrast factor. We apply the contrast transformation to only one modality at a time while keeping the other modality at its original contrast. We evaluate segmentation performance on both L and H&T using UNet under two contrast settings: $k = 0.8$ (reduced contrast) and $k = 1.2$ (enhanced contrast).

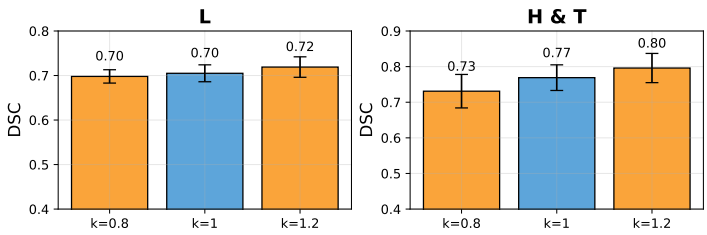

Figure 11: **Model performance under different CT contrast levels (with original MRI).** $k = 1$ represents the original images without any contrast changes.

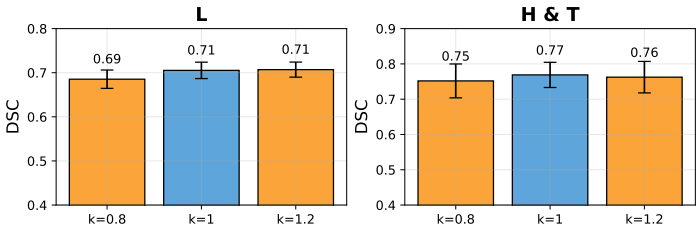

Figure 12: **Model performance under different MRI contrast levels (with original CT).** $k = 1$ represents the original images without any contrast changes.

As shown in Figure 11 and 12, the performance of ContourDiff remains largely stable, with slight improvements with enhanced contrast and minor degradations with reduced contrast. This is expected, as higher contrast produces clearer extracted contours and preserves more anatomical details. Therefore, in practice, enhancing contrast for both input-domain and output-domain images prior to applying the method may lead to better performance.

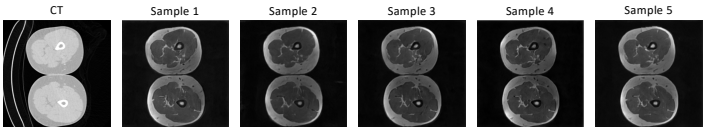

Figure 13: **Resamples visualization.** Multiple translated samples conditioned on the same input-domain information.

4.8.4 Sampling Stability Assessment

We further evaluate the sampling stability of our model. First, we manually inspect 10 directly translated volume without any slice selection and find that, on average, 16.92% of slices within each volume exhibit overly bright backgrounds. Next, we quantitatively assess sampling consistency across stochastic runs by computing the mean pixel-wise variance across multiple translated samples per slice. Specifically, we calculate the variance over 100 randomly selected CT contours, each with 10 generated samples, and the output mean variance is 0.007. As shown in Figure 13, translated samples from the same input-domain conditions under different seeds remain anatomically stable. The modest variability suggests some residual sampling instability. Therefore, to further stress test the initial slice selection process, we performed experiments by varying (1) the number of candidates for initial slice selection, and (2) the selection of suboptimal candidates for initial slice.

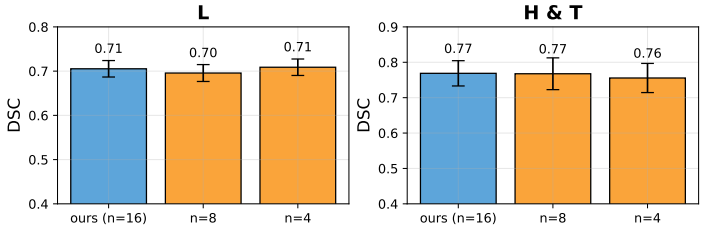

Figure 14: **Model performance with varying number of candidate slices for initial slice selection.** $k = 16$ represents the current setting.

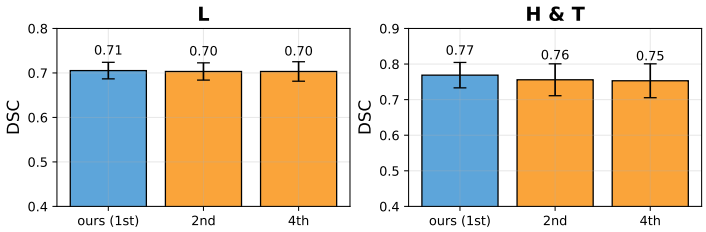

Figure 15: **Model performance when selecting different ranks of initial slice candidates.** $1st$ represents the current setting.

As shown in Figure 14 and 15, although sampling can introduce some instability, ContourDiff demonstrates strong robustness when fewer candidate slices are available or when suboptimal candidates are selected, with a DSC reduction of at most 0.02.

4.9 Experiments on Liver Translation

We perform CT to MRI liver translation using the public AMOS dataset Ji et al. (2022) to further demonstrate the capability of ContourDiff on soft-tissue structures. Following previous settings, we collect 1,077 2D axial MRI slices for training ContourDiff model. For the downstream liver segmentation task, we randomly split 2,126 2D axial CT slices by volume into 1,625 for training and 501 for validation, and hold out 695 2D axial MRI slices as the test set.

	Liver		
Method	DSC (↑)	ASSD (↓)	Edge HD95 (↓)
w/o Adap.	0.299 ± 0.031	8.545 ± 0.641	-
CycleGAN	0.848 ± 0.027	2.419 ± 0.380	20.787 ± 0.281
UNSB	<u>0.872 ± 0.008</u>	<u>2.377 ± 0.208</u>	<u>16.962 ± 0.157</u>
Ours	**0.873 ± 0.012**	**2.193 ± 0.211**	**4.451 ± 0.069**
UB†	0.913 ± 0.015	1.404 ± 0.137	-

Table 7: Quantitative comparison (DSC, ASSD and Edge HD95) of ContourDiff on CT to MRI liver translation task. Best in bold, runner-up underlined.

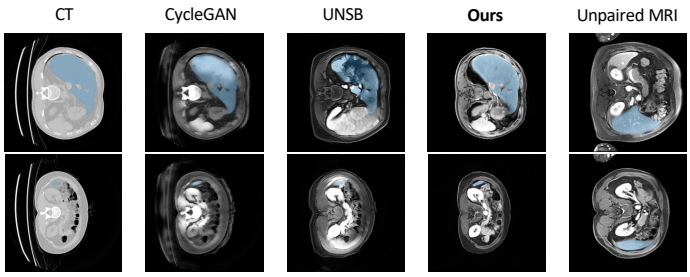

Figure 16: **Qualitative comparison of ContourDiff on liver translation.** ContourDiff seems to best keep anatomical consistency during translation for soft-tissue structures in abdominal area. The input-domain segmentation masks are presented in blue to visualize the alignment.

According to Table 7 and Figure 16, ContourDiff outperforms representative GAN-based (CycleGAN) and diffusion-based (UNSB) methods in liver translation, achieving superior performance in both segmentation accuracy and edge alignment, which demonstrates its capability for soft-tissue translation.

4.10 Evaluation on Higher Bits Normalization

To evaluate whether higher bit normalization affects performance, we compared our current 8-bit normalization (i.e., 0-255) with 12-bit normalization (i.e., 0-4095). As shown in Figure 17, the results show no substantial performance differences between the two settings.

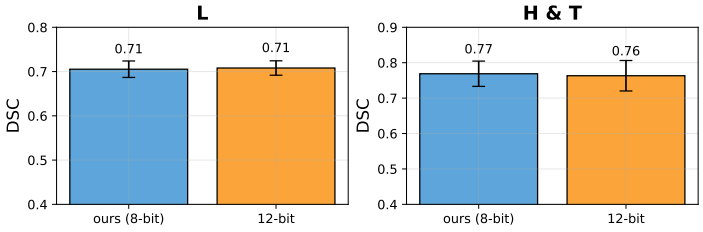

Figure 17: **Model performance with higher bit normalization**. 8-bit is the current setting.

4.11 Efficiency Evaluation

To assess the practicality and feasibility of clinical deployment, we report the runtime and the peak memory usage of ContourDiff as shown in Table 8:

Metric/Equipment	Results
GPU	1 NVIDIA RTX A6000
Training Batch Size	4
Peak Training GPU Usage	12.06 GB
# Candidates for initial slices	16
Peak Testing GPU Usage	10.40 GB
# DDIM Steps	50
Inference Time per 2D Slice (n parallel group)	1.592 s $/n$
Inference Time per 3D Volume (n parallel group)	210.789 s $/n$
Ave. # slices per 3D Volume	132.4
Total Training Time for ContourDiff	$\sim$ 5 hrs
Total Training Time for Segmentation	L: $\sim$ 20 min; H&T: $\sim$ 1 hr

Table 8: Efficiency assessment of ContourDiff.

5. Conclusion and Future Work

In this paper, we introduce a novel framework, ContourDiff with SCGD, to preserve anatomical fidelity in unpaired image translation. Our method constrains the generated images in the output domain to align with the anatomical contour of images from the input domain. Both quantitative and qualitative results on medical datasets show that ContourDiff (with/without SCGD) significantly outperforms multiple existing image translation methods in maintaining anatomical structures. Furthermore, we demonstrated the zero-shot capability of ContourDiff by translating T2-weighted MRI to T1-weighted MRI without any retraining.

As a direction for future work, the practical deployment of ContourDiff may further benefit from automatic threshold selection based on simple image statistics (e.g., percentile-based methods) and from including a lightweight contour-refinement network, thereby further reducing manual tuning. We also note that susceptibility-related MRI distortions can carry clinically relevant information, thus, strictly enforcing pixel-wise contour alignment could risk suppressing such meaningful distortions. Future research will explore distortion-aware guidance to balance structural fidelity with the preservation of diagnostically relevant geometric information. In addition, while ContourDiff reliably preserves common anatomical structures such as bone, fat, muscle, and major organs, applications requiring finer tissue characteristics may benefit from integrating additional input-domain information or a hybrid conditioning strategy. Moreover, incorporating real multi-contrast or multi-echo MRI data and evaluating performance under varying acquisition conditions would further enhance the generalizability of the proposed methods across diverse clinical settings.

Ethical Standards

The research protocol was approved by the Duke Health System Institutional Review Board (IRB) with ethical standards for research and manuscript preparation, adhering to all relevant laws and regulations concerning the treatment of human subjects and animals.

Conflicts of Interest

We declare we do not have any conflicts of interest.

Data availability

The external dataset analyzed in this study is publicly available (Wasserthal et al. (2023); van der Graaf et al. (2023); Ji et al. (2022)). The internal dataset, however, is not currently available due to an extensive de-identification procedures and institutional review board requirements. The code have been made publicly available at: `https://github.com/mazurowski-lab/ContourDiff`.

References

Karim Armanious, Chenming Jiang, Sherif Abdulatif, Thomas Küstner, Sergios Gatidis, and Bin Yang. Unsupervised medical image translation using cycle-medgan. In *2019 27th European signal processing conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation

using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Farzad Beizaee, Christian Desrosiers, Gregory A Lodygensky, and Jose Dolz. Harmonizing flows: Unsupervised mr harmonization based on normalizing flows. In *International Conference on Information Processing in Medical Imaging*, pages 347–359. Springer, 2023.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

Junhua Chen, Shenlun Chen, Leonard Wee, Andre Dekker, and Inigo Bermejo. Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review. *Physics in Medicine & Biology*, 2023a.

Yuwen Chen, Helen Zhou, and Zachary C Lipton. Moco-transfer: Investigating out-of-distribution contrastive learning for limited-data domains. *arXiv preprint arXiv:2311.09401*, 2023b.

Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

Alicia Durrer, Julia Wolleb, Florentin Bieder, Tim Sinnecker, Matthias Weigel, Robin Sandkuehler, Cristina Granziera, Özgür Yaldizli, and Philippe C Cattin. Diffusion models for contrast harmonization of magnetic resonance images. In *Medical Imaging with Deep Learning*, pages 526–551, 2024.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2427–2436, 2019.

Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K. Moyo, Michael R. Savona, Richard G. Abramson, and Bennett A. Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE Transactions on Medical Imaging*, 38(4):1016–1025, April 2019. ISSN 1558-254X. . URL `http://dx.doi.org/10.1109/TMI.2018.2876633`.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. *arXiv preprint arXiv:2305.15086*, 2023.

Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024.

Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34:1964–1978, 2021.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023a.

Wen Li, Yafen Li, Wenjian Qin, Xiaokun Liang, Jianyang Xu, Jing Xiong, and Yaoqin Xie. Magnetic resonance image (mri) synthesis from brain computed tomography (ct) images based on deep learning methods for magnetic resonance (mr)-guided radiotherapy. *Quantitative imaging in medicine and surgery*, 10(6):1223, 2020.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *arXiv preprint arXiv:2304.02742*, 2023b.

Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style transfer using generative adversarial networks for multi-site mri harmonization. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 313–322. Springer, 2021.

Fei Lyu, Jingwen Xu, Ye Zhu, Grace Lai-Hung Wong, and Pong C Yuen. Superpixel-guided segment anything model for liver tumor segmentation with couinaud segment prompt. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 678–688. Springer, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

Gourav Modanwal, Adithya Vellal, Mateusz Buda, and Maciej A Mazurowski. Mri image harmonization using cycle-consistent generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 259–264. SPIE, 2020.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

Vu Minh Hieu Phan, Zhibin Liao, Johan W Verjans, and Minh-Son To. Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.

Matteo Rossi and Pietro Cerveri. Comparison of supervised and unsupervised approaches for the generation of synthetic ct from cone-beam ct. *Diagnostics*, 11(8):1435, 2021.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:8702–8716, 2022.

Hristina Uzunova, Jan Ehrhardt, and Heinz Handels. Memory-efficient gan-based domain translation of high resolution 3d medical images. *Computerized Medical Imaging and Graphics*, 86:101801, 2020.

Jasper W. van der Graaf, Miranda L. van Hooff, Constantinus F. M. Buckens, Matthieu Rutten, Job L. C. van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark, 2023.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, June 2014. ISSN 2167-8359. . URL `http://dx.doi.org/10.7717/peerj.453`.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

Zihao Wang, Yingyu Yang, Yuzhou Chen, Tingting Yuan, Maxime Sermesant, Hervé Delingette, and Ona Wu. Mutual information guided diffusion for zero-shot cross-modality medical image translation. *IEEE Transactions on Medical Imaging*, 2024.

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), September 2023. ISSN 2638-6100. . URL `http://dx.doi.org/10.1148/ryai.230024`.

Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 255–263. Springer, 2019.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251, 2018.

Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.