

ContourDiff: Unpaired Image-to-Image Translation with Structural Consistency for Medical Imaging

Yuwen Chen^{1*} Nicholas Konz¹ Hanxue Gu¹ Haoyu Dong¹ Yaqian Chen¹
 Lin Li⁴ Jisoo Lee² Maciej A. Mazurowski^{1,2,3,4}

¹ Department of Electrical and Computer Engineering, Duke University, NC, USA

² Department of Radiology, Duke University, NC, USA

³ Department of Computer Science, Duke University, NC, USA

⁴ Department of Biostatistics & Bioinformatics, Duke University, NC, USA

<https://github.com/mazurowski-lab/ContourDiff>

Abstract

Preserving object structure through image-to-image translation is crucial, particularly in applications such as medical imaging (e.g., CT-to-MRI translation), where downstream clinical and machine learning applications will often rely on such preservation. However, typical image-to-image translation algorithms prioritize perceptual quality with respect to output domain features over the preservation of anatomical structures. To address these challenges, we first introduce a novel metric, StrúctB, to quantify the **structural bias** between domains which must be considered for proper translation. We then propose **ContourDiff**, a novel image-to-image translation algorithm that leverages domain-invariant anatomical contour representations of images to preserve the anatomical structures during translation. These contour representations are simple to extract from images, yet form precise spatial constraints on their anatomical content. ContourDiff applies an input image contour representation as a constraint at every sampling step of a diffusion model trained in the output domain, ensuring anatomical content preservation for the output image. 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.

1. Introduction

Unpaired image-to-image (I2I) translation—the task of translating images from some input domains to an out-

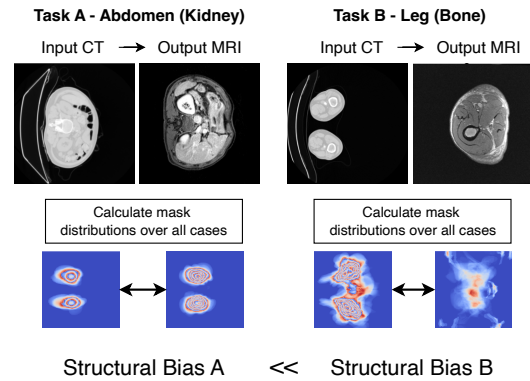


Figure 1. **Structural biases between CT and MRI modalities in certain anatomical regions:** small for the abdominal region (for kidneys) from axial view (Task A), but large for the leg (for bones) from axial view (Task B). Top half: example images from the two domains; bottom half: average object masks \hat{m} for each domain. We focus on datasets with significant structural bias in this paper.

put domain with only unpaired data for training [56]—offers extensive applications in medical image analysis [2, 5, 14, 32, 33, 50, 52, 54]. A significant use case is facilitating segmentation across different imaging modalities (e.g., CT and MRI) [10], for anatomical locations such as brain [30], abdomen [22], and pelvis [41]. 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.

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 (Sec. 3.1). An example of this is the drastic visual difference between CT and MRI for leg and spinal re-

*Corresponding author: yuwen.chen@duke.edu

gions as captured in standard exams (Fig. 1 and Fig. 5 in App. D), 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) [17] such as Cycle-consistent Adversarial Network (CycleGAN) [1, 10, 38, 55, 56]. These methods maintain the consistency between the images from input and output domains by leveraging cycle consistency loss, minimizing information loss during bidirectional translation [56]. 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 [38].

Recently, several conditional diffusion models have been introduced for image translation tasks, both in natural images [4, 25, 29, 39] and medical imaging [26, 31, 36]. 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 [28, 39, 53], we propose a diffusion model for image translation, “**ContourDiff**”, 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, 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. 5 in App. D). 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 identify and quantify the structure bias problem of

- image translation by proposing a new metric, $\hat{\text{StructB}}$.
2. We propose ContourDiff, a novel diffusion-based method for unpaired image-to-image translation which allows zero-shot learning.
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.

2. Related Work

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 [23] and its variants [49], 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 [56] introducing cycle-consistency loss to allow translation between unpaired datasets. MUNIT [21] enables multi-modal outputs to generate diverse outputs given images from input domains. GcGAN [15] incorporates geometric-consistency constraints to preserve the geometric information across domains. To reduce the training time, CUT [37] 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 [29]. More recently, diffusion-based translation frameworks have emerged as a promising alternative, providing competitive performance in both paired [29] and unpaired [25] 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 [11], 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 [1, 27, 46]. Recently, diffusion models have gained popularity in this domain. For instance, SynDiff [36] 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 constraints on the detailed anatomical structures during translation.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) [19], or just *diffusion models*, have recently gained significant attention for their remarkable performance in generative modeling across both natural [13, 34] and medical imaging tasks [28, 39]. 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. [28] 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) [39] 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 [12, 39], super-resolution [16, 42] and semantic segmentation [3, 45].

3. Methods

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_n^{out} ($n = 1, \dots, N_{\text{out}}$) are available to train on. The goal is then to use the trained model at inference to translate unseen input domain data x_n^{in} 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 model-based image translation framework based on domain-invariant anatomical contour representations of images.

3.1. The Problem of Structural Bias

One of the primary motivations for our method is the issue of *structural bias* seen in many medical image translation problems. Qualitatively, we define structural bias as when anatomical structure and shape *consistently* differ in some way between the input and output domains. Formally, we describe the structure of a given image $x \in \mathbb{R}^N$ in terms of its objects, via some binary segmentation masks $m \in \{0, 1\}^N$ for a given object; for example, the objects could be bones in our dataset. We define the structural bias between the two domains in terms of their objects, as follows.

Definition 3.1 (Structural Bias Between Domains). Consider respective input/source and output/target domain image distributions $p(x^{\text{in}})$ and $p(x^{\text{out}})$, where each image x has a binary mask m for some objects, defining corresponding mask distributions $p(m^{\text{in}})$ and $p(m^{\text{out}})$. The *structural bias* between the two domains is defined in a novel way as

$$\text{StructB} := \mathbb{E}_{m^{\text{in}} \sim p(m^{\text{in}})} \mathbb{E}_{m^{\text{out}} \sim p(m^{\text{out}})} \|m^{\text{in}} - m^{\text{out}}\|_2. \quad (1)$$

In other words, a pair of domains have high structural bias if two randomly sampled masks from each of the two domains *consistently* differ, on average.

Given input/output domain datasets of respective sizes N_{in} and N_{out} , we could estimate their structural bias via Eq. 1, but this would not be tractable for large datasets due to its scaling as $\mathcal{O}(N_{\text{in}}N_{\text{out}})$. Instead, we can compute a tractable *lower bound* for the structural bias which scales linearly as $\mathcal{O}(\max(N_{\text{in}}, N_{\text{out}}))$,

$$\text{Struct}\hat{\text{B}} := \|\mathbb{E}_{m^{\text{in}} \sim p(m^{\text{in}})}(m^{\text{in}}) - \mathbb{E}_{m^{\text{out}} \sim p(m^{\text{out}})}(m^{\text{out}})\|_2, \quad (2)$$

or just $\text{Struct}\hat{\text{B}} = \|\hat{m}^{\text{in}} - \hat{m}^{\text{out}}\|_2$, notating the *average* masks of each domain as $\hat{m}^{\text{in}} := \mathbb{E}_{m^{\text{in}} \sim p(m^{\text{in}})}(m^{\text{in}})$ and similar for \hat{m}^{out} . It is easy to prove that $\text{Struct}\hat{\text{B}} \leq \text{StructB}$ via Jensen’s inequality (as the norm is convex)¹. In addition, the size of the object of interest should be considered as the same amount of bias can have varying impacts on large versus small objects. In practice, we resize each mask to the same dimension to ensure $\hat{m} \in \mathbb{R}^{H \times W}$. Then, we compute $\text{Struct}\hat{\text{B}}$ normalized with respect to (1) the number of pixels in the masks/images and (2) the average intensity of \hat{m}^{in} and \hat{m}^{out} ,

$$\text{Struct}\hat{\text{B}} = \frac{2\|\hat{m}^{\text{in}} - \hat{m}^{\text{out}}\|_2}{\sqrt{N}(\text{Avg}[\hat{m}^{\text{in}}] + \text{Avg}[\hat{m}^{\text{out}}])}. \quad (3)$$

$$\text{Avg}[\hat{m}] := \frac{1}{N} \sum_{i=1}^H \sum_{j=1}^W \hat{m}_{i,j}, \quad (4)$$

where $N = H \times W$ is the size of masks. We show example images from CT and MRI pairs with both small and large structural biases in Fig. 1 (upper half), alongside the average masks \hat{m}^{in} and \hat{m}^{out} for each domain (lower half). Leg and spine datasets have larger structural biases than the abdominal dataset (Table 1, Fig. 5 in App. D), shown by how much the domains’ average masks \hat{m}^{in} and \hat{m}^{out} differ.

This formalism illustrates why structural bias is not a primary concern in mainstream computer vision datasets used for image translation model evaluation, as these models are generally not designed to address it. For instance, in the

¹Indeed, one reason we use the L_2 distance to compare masks rather than IoU or Dice coefficient is because the latter two are not convex functions, making the Jensen’s Inequality approximation not applicable.

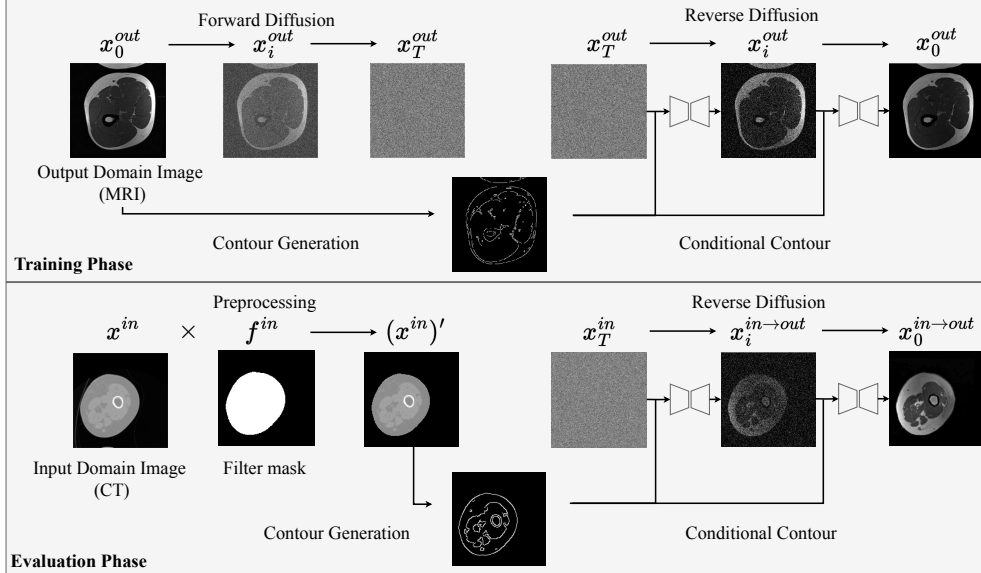


Figure 2. Diagram of ContourDiff pipeline. Preprocessing on x^{out} is omitted due to clean background of MRIs.

Dataset (Target):	Abd. (Kidney) [24]	L (Bone) [51]	H & T (Bone) [51]
StructB :	0.129	0.287	0.218

Table 1. Estimated structural biases StructB (Eq. 3) between the domains of the datasets shown in Fig. 5 in App. D. “Abd.”: Abdomen, “L”: Lumbar Spine, “H & T”: Hip & Thigh.

widely-used horse \leftrightarrow zebra and summer \leftrightarrow winter Yosimete datasets (as shown in Fig. 13 in [56]), randomly selected images from each domain do not consistently show differences in object shape or position (e.g., horses/zebras or mountains, respectively) beyond what is observed within the same domain, potentially leading to low StructB². In contrast, medical imaging modalities, such as CT and MRI, often involve different acquisition protocols. Such protocols can result in a given object (e.g., kidneys) having consistent shape and position within a single domain, yet different imaging settings between domains (e.g., field of view) can lead to noticeably consistent variations in object appearance and positioning, resulting in large StructB (Fig. 1).

Previous work [9] has demonstrated the capability of GAN-based models for CT-to-MRI translation in abdominal regions. For example, [9] reported kidney segmentation performance trained on CycleGAN-translated MRIs achieved a Dice Coefficient over 0.768 when testing on real MRIs. Thus, we mainly focus on datasets with larger StructB (i.e., L and H & T) in this paper.

Unlike existing translation models, ContourDiff is specifically designed for domains with high structural bias. ContourDiff explicitly ensures that the pixelwise structure/content seen in the input image is present in the trans-

²The absence of foreground masks prevents calculation of StructB.

lated image, more so than prior content-preserving translation models based on style/content-disentanglement (e.g., MUNIT [21]) which do not enforce this explicitly. We will show this empirically (As shown in Table 2, Fig. 3).

3.2. Adding Contour Guidance to Diffusion Models

3.2.1 A Review of Diffusion Models

Denoising diffusion probabilistic models [19] are generative models that learn to reverse a gradual process of adding noise to an image over many time steps $t = 0, \dots, T$. 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 $t = T, \dots, 0$ 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 [40]) 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 described as [35]:

$$L = \mathbb{E}_{x_0, t, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (5)$$

where θ is the model parameters.

Unlike unconditional DDPMs, many conditional diffusion models [28, 29, 39] directly integrate the conditions y (e.g., images and texts) into the training objective:

$$L = \mathbb{E}_{(x_0, y), t, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t|y)||^2], \quad (6)$$

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

Denoising Diffusion Implicit Models (DDIMs) [43] employ a deterministic, non-Markovian sampling process, allowing for faster sample generation without noticeable compromises for image fidelity.

Algorithm 1 Contour-guided DDPM model training.**Input:** Output domain training distribution $p(x_0^{\text{out}})$.**repeat**

$$\begin{aligned}
& x_0^{\text{out}} \sim p(x_0^{\text{out}}) \\
& c^{\text{out}} = \text{Canny}(x_0^{\text{out}}) \\
& \epsilon \sim \mathcal{N}(0, I_n) \\
& t \sim \text{Uniform}(\{1, \dots, T\}) \\
& x_t^{\text{out}} = \sqrt{\bar{\alpha}_t} x_0^{\text{out}} + \sqrt{1 - \bar{\alpha}_t} \epsilon \\
& \text{Update } \theta \text{ with } \quad \nabla_{\theta} \|\epsilon - \epsilon_{\theta}(x_t^{\text{out}}, t|c^{\text{out}})\|^2
\end{aligned}$$
until converged;

Algorithm 2 Contour-guided image translation.**Input:** Input domain image x^{in} .**Output:** Translated image $x_0^{\text{in} \rightarrow \text{out}}$ $c^{\text{in}} = \text{Canny}(x^{\text{in}})$ $x_T^{\text{out}} \sim \mathcal{N}(0, I_n)$ **for** $t = T, \dots, 1$ **do** $\epsilon \sim \mathcal{N}(0, I_n)$ if $t > 1$, else $\epsilon = 0$ $x_{t-1}^{\text{out}} = \frac{1}{\sqrt{\alpha_t}} \left(x_t^{\text{out}} - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta}(x_t^{\text{out}}, t|c^{\text{in}}) \right) + \sigma_t \epsilon$ **end****return** $x_0^{\text{in} \rightarrow \text{out}}$

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 [7] to extract the contour representation c of each training image x_0 , and concatenate it with the network input at every denoising step, a practice similar to [28, 53]. This modifies the network in Eq. 6 to become $\epsilon_{\theta}(x_t, t|c)$ and the diffusion training objective to become

$$L = \mathbb{E}_{(x_0, c), t, \epsilon} [\|\epsilon - \epsilon_{\theta}(x_t, t|c)\|^2], \quad (7)$$

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 [28]), which may be lost if diffusion is performed within a latent space [39].

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 domains 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 computed contours $(x_n^{\text{out}}, c_n^{\text{out}})$, shown in Algorithm 1. Next, to translate some *input 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 $x^{\text{in} \rightarrow \text{out}}$. Therefore, $x^{\text{in} \rightarrow \text{out}}$ 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 $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

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 $x \leftarrow M_{\text{filter}} \odot x$. M_{filter} is defined by sequentially computing the follow Scikit-Image [48] functions on x [38]: `threshold_multiotsu`, `binary_erosion`, `remove_small_objects`, and `remove_small_holes`.

3.3.3 Consistent Translation in Adjacent Slices

Additionally, we propose to enforce the consistency of translating adjacent input domain image slices taken from 3D images (e.g., CT) to a output domain as follows. Firstly, we translate the first slice image $[x^{\text{in}}]_1$ in the 3D volume to its output domain version $[x^{\text{in} \rightarrow \text{out}}]_1$. We repeat the generation until the mean of one generated image is less than a specified threshold m_{thresh} . Then, we translate the successive slices $[x^{\text{in}}]_i$ ($i = 2, \dots, N_{\text{slices}}$) by generating different candidates $[x^{\text{in} \rightarrow \text{out}}]_i$ (starting with different sampled noise) and selecting one that is within an L_2 distance of δ of the previous slice translation $[x^{\text{in} \rightarrow \text{out}}]_{i-1}$. The used threshold values are shown in Table 6 in App. B. During each iteration, if multiple candidates satisfy the L_2 distance criterion, we choose the one with the smallest δ . We generate 4 candidates per iteration, allowing up to 5 attempts; should none of the candidates meet the specified requirements after this, we select the one with the smallest δ .

4. Experiments

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 [51], SPIDER lumbar spine (L-SPIDER) public dataset [47] and a private in-house dataset. For the MRIs used to train the Contour-Diffusion, we collect a private dataset with T-1 weighted lumbar spine (L) and hip & thigh (H&T) body regions. 40 sagittal lumbar MRI volumes (670 2D slices), and 10 axial MRI volumes

Method	L				L-SPIDER				H & T			
	UNet		SwinUNet		UNet		SwinUNet		UNet		SwinUNet	
	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)
w/o Adap.	0.287	6.515	0.171	7.386	0.236	8.275	0.187	8.327	0.004	45.730	0.003	48.624
CycleGAN [56]	0.484	2.479	0.362	3.505	0.507	3.629	0.412	3.701	<u>0.535</u>	9.140	<u>0.464</u>	<u>9.791</u>
SynSeg-Net [22]	0.316	3.014	0.288	3.527	0.364	3.207	0.291	5.502	0.370	<u>4.708</u>	0.059	12.869
CyCADA [20]	0.331	5.942	0.319	3.691	0.364	4.389	0.260	4.726	0.349	11.247	0.155	13.004
MUNIT [21]	0.407	3.804	0.433	3.212	0.380	4.309	0.358	3.545	0.128	16.229	0.090	18.925
CUT [37]	0.392	4.669	0.288	5.259	0.368	5.781	0.292	6.751	0.311	19.252	0.211	20.564
GcGAN [15]	<u>0.554</u>	<u>1.753</u>	0.433	<u>2.940</u>	<u>0.580</u>	<u>2.202</u>	<u>0.513</u>	<u>2.904</u>	0.414	9.275	0.320	13.649
MaskGAN [38]	0.428	3.192	0.322	4.692	0.458	3.729	0.385	5.355	0.289	16.228	0.292	17.591
UNSB [25]	0.465	3.111	<u>0.456</u>	2.955	0.488	3.984	0.446	3.070	0.247	13.427	0.181	17.650
Ours	0.683	1.432	0.654	1.434	0.633	2.066	0.534	2.353	0.731	3.139	0.659	5.780

Table 2. Comparison of our model 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. Best in bold, runner-up underlined.

Lumbar Spine (L) - Foreground										
Metric	CycleGAN [56]	SynSeg-Net [22]	CyCADA [20]	MUNIT [21]	CUT [37]	GcGAN [15]	MaskGAN [38]	UNSB [25]	Ours	
FID (\downarrow)	132.16	137.63	127.54	372.67	150.10	138.60	<u>128.17</u>	137.42	122.75	
KID (\downarrow)	0.047	0.054	0.045	0.343	0.058	0.050	0.039	0.051	<u>0.041</u>	
Hip & Thigh (H & T) - Foreground										
Metric	CycleGAN [20]	SynSeg-Net [22]	CyCADA [20]	MUNIT [21]	CUT [37]	GcGAN [15]	MaskGAN [38]	UNSB [25]	Ours	
FID (\downarrow)	183.18	192.32	184.11	193.12	193.63	<u>163.61</u>	175.28	167.88	135.39	
KID (\downarrow)	0.163	0.169	0.159	0.174	0.178	0.144	0.152	<u>0.142</u>	0.101	

Table 3. Comparison of foreground FID and KID between translated images and output domain images. Best in bold, runner-up underlined. (Note: L-SPIDER is excluded as it is only used for testing and not included in training translation model.)

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 [51] 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 [47]³.

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). As there are no paired images, we also calculate the foreground⁴ FID [18] and KID [6] 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. Comparison with Other Methods

We compare our method to 8 other state-of-the-art translation/adaptation methods, including CycleGAN [56], SynSeg-Net [22], CyCADA [20], MUNIT [21], CUT [37], GcGAN [15], MaskGAN [38] and UNSB [25], via the performance of output domain-trained downstream task segmentation models on translated images. Several of these methods (e.g., [15, 21, 22, 25, 37, 38, 56]) translate the images solely at the image level, while CyCADA also aligns the latent feature output from the model encoder of downstream task. Mask-GAN incorporates the extracted coarse masks to better preserve object structures throughout translation. In addition to GAN-based model, UNSB combines diffusion models with Schrödinger Bridge theory to enable probabilistically consistent translation for unpaired data. For CyCADA, we utilized the same segmentation architecture as the other methods but without the skip connection to enable feature-level alignment. For each competing method, we evaluated multiple intermediate results for the translation tasks (see App. C). The best performance among these results is reported.

³We crop the slices to exclude the sacrum, as it is not annotated.

⁴Foreground refers to pixels containing the object of interest. In this paper, we use masks from CTs to extract objects.

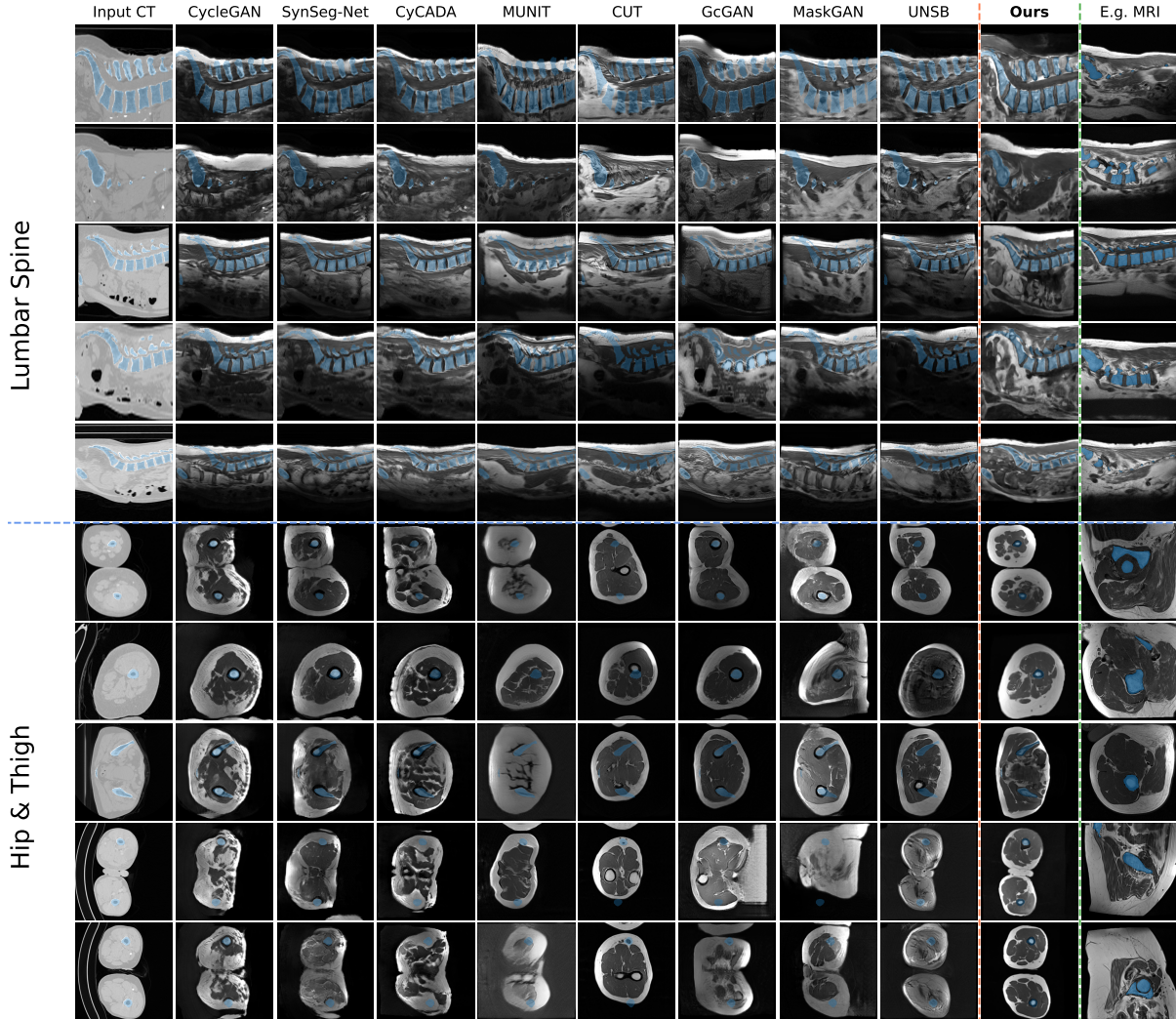


Figure 3. Generated MRIs given CTs from Lumbar and Hip & Thigh areas from different translation models. The masks (in blue) from the original CTs are added to all the generated images to visualize the alignment. The real MRIs in the last right column are unpaired and included for visualization of output domain (not ground truth).

4.4. Implementation Details

We adopt the UNet architecture [40] for the denoising model ϵ_θ with a two-channel input (grayscale image and its contour). The training settings for the diffusion model follow the same as that in [28].

We use the DDIM algorithm [44] for sampling, with 50 steps. For the segmentation models, we use the convolution-based UNet [40] and transformer-based SwinUNet [8]. All images are resized to 256×256 and normalized to $[0, 255]$. The training of competing methods mostly follows the default settings from each official GitHub. We set $\lambda_{idt} = 0.5$ to include identity loss if the methods are provided. We train the downstream segmentation model with a cosine learning rate scheduler up to 100 epochs with the initial learning rate of 1×10^{-3} .

4.5. Results

Quantitative Results. The segmentation model results are shown in Table 2. For the three test sets, our method outperforms previous image adaptation methods by a significant margin: for example, the UNet DSC on output domain segmentation increase by 0.129, 0.053 and 0.196 for L, L-SPIDER and H&T, respectively, compared to the second best. Also, segmentation models trained on CycleGAN-translated images achieved around 0.5 DSC on L and H & T (see Table 2), which is significantly lower than DSC for kidney (i.e., datasets/tasks with lower StructB) reported in [9], despite minor model differences.

Based on Table 3, our method achieves the lowest FID scores: 122.75 and 135.39 for L and H & T, respectively. For KID scores, our method outperforms others for H & T

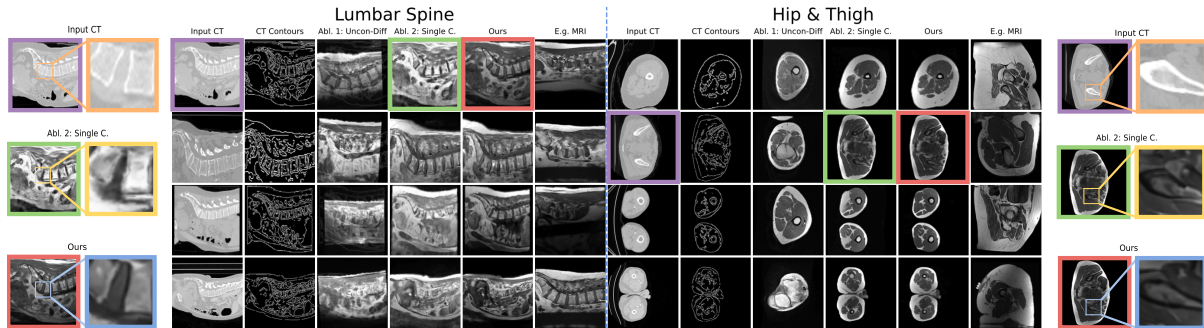


Figure 4. Qualitative results of ablation study with zoomed-in anatomical details highlighting on the sides.

and achieves a close second place for L (0.041), which is slightly lower than the top score of 0.039 by MaskGAN.

Qualitative Results. We provide example image translations in Fig. 3. 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 over-emphasize these features and, therefore, tend to translate CTs of two legs into MRIs depicting only one leg (Fig. 3). For the lumbar spine from the sagittal view, MRIs often start from the lowest thoracic spine and end at the sacrum. On the other hand, CTs often include the upper leg and sometimes the abdominal body (see Fig. 5 in App. D). 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.

Based on Table 2, Table 3 and Fig. 3, ContourDiff best maintain anatomical fidelity compared to other models, both quantitatively and qualitatively.

4.6. Ablation Study

We conduct ablation studies to validate the effectiveness of several key designs in ContourDiff.

Effectiveness of Adding Contours. We verify the effectiveness of introducing contours to each denoising step during training by conditionally training on empty map (i.e., all zeros) and adding the CTs contours during the translation steps. Fig. 4 showed that the denoised model ϵ_{θ}

trained without contours hardly followed the introduced CTs contours (**‘Uncon-Diff’** column). Furthermore, the UNet trained on these unconditionally generated MRIs experienced a dramatic performance drop (see Table 4).

Single Candidate Generation. We generate the images directly (i.e., by single candidate) without enforcing translation consistency for adjacent slices (mentioned in 3.3.3). The qualitative result shows a reduced quality of the generated images, including incorrect contrast and anatomical consistency, by using a single candidate (see Fig. 4 **‘Single C.’** column), leading to degraded performance for segmentation models trained on these images (see Table 4).

Method	M_{seg}	L		L-SPIDER		H & T	
		DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)	DSC (\uparrow)	ASSD (\downarrow)
Uncon-Diff	UNet	0.354	5.360	0.197	7.251	0.281	19.895
Single C.	UNet	0.627	2.022	0.571	2.370	0.624	4.752
Ours	UNet	0.683	1.432	0.633	2.066	0.731	3.139

Table 4. Quantitative results of ablation study in terms of segmentation model (M_{seg}) performance.

5. Conclusions and Future Work

In this paper, we first identified structural bias problems during I2I translation and proposed a new metric, StructB, to quantify such bias. We then introduced a novel method (ContourDiff) to preserve the 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 significantly outperforms multiple existing image translation methods in preserving anatomical structures.

Nevertheless, one key limitation of ContourDiff could be: there is a need to select several hyperparameters in the translation stage as outlined in Section 3.3.3. Future work could aim to enhance control over the translation process to preserve the consistency between adjacent slices. In addition, another interesting direction could be extending our method to multi-domain medical image harmonization.

References

- [1] 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 (EU-SIPCO)*, pages 1–5. IEEE, 2019. 2
- [2] 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. 1
- [3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3
- [4] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 2
- [5] Farzad Bezaee, 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. 1
- [6] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6
- [7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 5
- [8] 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. 7
- [9] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020. 4, 7
- [10] 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*, 2023. 1, 2
- [11] 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*, 2023. 2
- [12] 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. 3
- [13] 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. 3
- [14] 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. 1
- [15] 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. 2, 6, 1
- [16] 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. 3
- [17] 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. 2
- [18] 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. 6
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4
- [20] 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. 6, 1
- [21] 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. 2, 4, 6, 1
- [22] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K. Moyo, Michael R. Savona, Richard G. Abramson, and Bennett A. Landman. Synsegnet: Synthetic segmentation without target modality ground truth. *IEEE Transactions on Medical Imaging*, 38(4):1016–1025, 2019. 1, 6
- [23] 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. 2
- [24] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xi-ang 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. 4

- [25] 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. [2](#), [6](#), [1](#)
- [26] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multimodal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024. [2](#)
- [27] 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. [2](#)
- [28] 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. [2](#), [3](#), [4](#), [5](#), [7](#)
- [29] 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, 2023. [2](#), [4](#)
- [30] 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. [1](#)
- [31] 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*, 2023. [2](#)
- [32] 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. [1](#)
- [33] 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*, pages 259–264. SPIE, 2020. [1](#)
- [34] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarbuerger, 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. [3](#)
- [35] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. [4](#)
- [36] Muzaffer Özbey, Onat Dalmaç, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Un-supervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. [2](#)
- [37] 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. [2](#), [6](#), [1](#)
- [38] 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. [2](#), [5](#), [6](#), [1](#)
- [39] 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. [2](#), [3](#), [4](#), [5](#)
- [40] 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. [4](#), [7](#)
- [41] 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. [1](#)
- [42] 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. [3](#)
- [43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [4](#)
- [44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. [7](#)
- [45] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:8702–8716, 2022. [3](#)
- [46] 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. [2](#)
- [47] 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. [5](#), [6](#)
- [48] 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, 2014. [5](#)

- [49] 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. [2](#)
- [50] 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. [1](#)
- [51] 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), 2023. [4](#), [5](#), [6](#)
- [52] 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. [1](#)
- [53] 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. [2](#), [5](#)
- [54] 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. [1](#)
- [55] 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. [2](#)
- [56] 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. [1](#), [2](#), [4](#), [6](#)

ContourDiff: Unpaired Image-to-Image Translation with Structural Consistency for Medical Imaging

Supplementary Material

A. Training time for Existing I2I Models

The total amount of training time for each competing benchmark is shown in Table 5 below:

Models	Total Training Time
CycleGAN [56]	200 epochs
SynSeg-Net [22]	200 epochs
CyCADA [20]	200 epochs
MUNIT [21]	1,000,000 iterations/steps
CUT [37]	400 epochs
GcGAN [15]	200 epochs
MaskGAN [38]	200 epochs
UNSB [25]	60 for L, 280 for H & T

Table 5. Training time for existing benchmark models.

B. Used Threshold Values

The used threshold values are presented in Table 6 below:

Dataset	m_{thresh}	δ
Lumbar Spine (L)	110	50
Hip & Thigh (H & T)	100	40

Table 6. Used threshold values in the experiments.

C. Checkpoint Evaluation Details

For SynSeg-Net and CyCADA, we evaluate the segmentation model every 20 epochs. For CycleGAN, MUNIT, CUT, GcGAN, MaskGAN and UNSB, as we need to train the segmentation model separately, we evaluate at 10%, 30%, 50%, 75% and 100% of the total training time. The total number of training time is shown in Tab. 5.

D. Structural Bias Diagram

The illustration of structural bias is shown below:

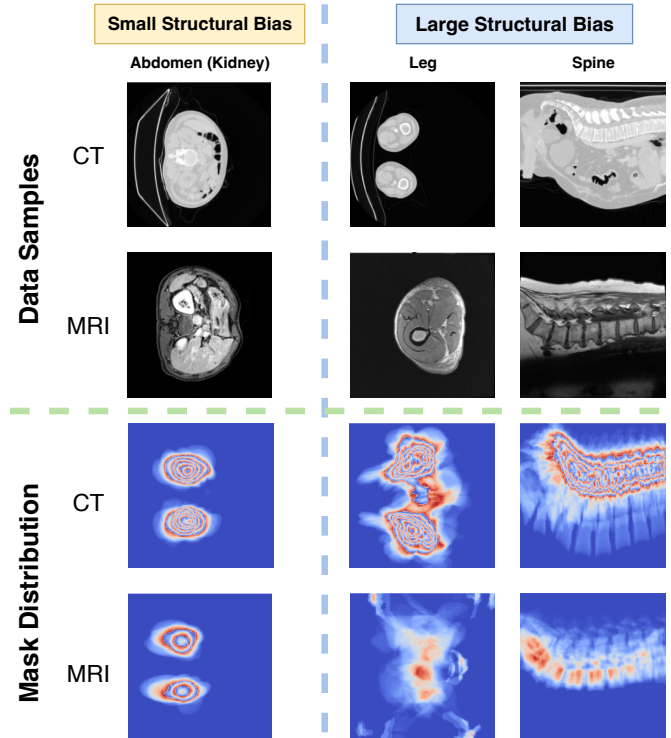


Figure 5. **Structural biases between CT and MRI modalities in certain anatomical regions:** small for the abdominal region (for kidneys) from axial view, but large for the leg and lumbar spine (for bones) from axial view. Top half: example images from the two domains; bottom half: average object masks \hat{m} for each domain.