MoDiff: A Morphology-Emphasized Diffusion Model for Ambiguous Medical Image Segmentation

Jung Su Ahn[1][0009−0006−5791−3136], Ki Hoon Kwak[2][0009−0007−4894−8959], Jung Woo Seo[2][0009−0005−9202−5321], and Young-Rae Cho[3,4][0000−0002−4645−2542]

[1] Graduate School of Computer Science, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
2023312031@yonsei.ac.kr

[2] Department of Computer Science and Engineering, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
{kihoon090,tjwjddn15584}@yonsei.ac.kr

[3] Department of Software, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea

[4] Department of Digital Healthcare, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
youngcho@yonsei.ac.kr

Abstract. MoDiff is a morphology-emphasized diffusion model designed for ambiguous medical image segmentation. It replaces traditional one-hot encoding with probability-based label maps to capture inherent uncertainties and ensure consistent segmentation results. By determining the presence of individual radiologist labels, MoDiff enables diverse sampling that provides richer insights into ambiguous areas. Its Learnable Discrete Frequency Filter (LDF) extracts high-frequency details for improved boundary precision, and when integrated with the Morphology-based Cross Attention Network (MCA), it enhances feature synthesis for more accurate anatomical segmentation. Evaluations on the LIDC-IDRI and MS-MRI datasets confirm its superior accuracy, boundary precision, and consistency.

Keywords: Attention Mechanism · Diffusion · Segmentation · Fast Fourier Transform

1 Introduction

In recent years, probabilistic segmentation models—such as the Probabilistic U-Net [16], Conditional Variational Autoencoders (cVAE) [1], and diffusion models [5, 12] have shown promise in addressing uncertainties in medical image analysis by generating multiple plausible segmentation hypotheses. However, sampling repeatedly from learned probability distributions can lead to inconsistent or overly varied outputs, complicating precise delineation of critical structures like organ boundaries. This inconsistency may result in conflicting clinical interpretations,

especially in tasks such as tumor diagnosis or organ segmentation. Moreover, training by randomly selecting reference labels fails to utilize all available annotation information.

To overcome these limitations, we propose the Morphology-Emphasized Diffusion Model for Ambiguous Medical Image Segmentation (MoDiff). MoDiff processes ambiguous medical images using a diffusion-based probabilistic segmentation approach that emphasizes morphological consistency. Unlike traditional methods using one-hot encoded label maps, we train the model with probability distribution-based label maps, enabling it to capture the inherent variability in medical annotations and produce consistent segmentation results across multiple samples.

Furthermore, to effectively extract morphological features and enhance noise reduction during the reverse diffusion process, we introduce the Learnable Discrete Frequency Filter (LDF) and the Morphology-based Cross Attention Network (MCA). LDF detects subtle boundary details and filters high-frequency noise using learnable parameters, while MCA synthesizes the derived features into a robust condition for denoising. This combined approach facilitates faster and more accurate learning of morphological structures compared to traditional methods.

2 Related Work

2.1 Stochastic Segmentation for Medical Imaging

Medical imaging inherently presents uncertainties due to device noise, low resolution, and anatomical variations. Stochastic segmentation techniques address these challenges by modeling uncertainty in the latent space to generate diverse segmentation outputs [3]. Notably, the Probabilistic U-Net [16] combines the U-Net architecture with a cVAE to produce multiple plausible segmentation hypotheses, offering more consistent results than pixel-wise probability models. Furthermore, cFlow [6] utilizes normalizing flows to extend simple Gaussian distributions into complex latent representations, while MoSE [7] and CIMD [2] capture uncertainty through multi-modal Gaussian and diffusion-based approaches, respectively.

2.2 Conditional Diffusion Model

Conditional Diffusion Models (CDMs) guide the denoising process with additional conditions to generate high-quality, condition-specific outputs [11]. BerDiff [9] replaces continuous Gaussian noise with the Bernoulli distribution to better handle discrete segmentation tasks. MedSegDiff [10] further enhances performance by integrating Dynamic Conditional Encoding and FFT-based noise suppression into a U-Net framework. Additionally, CCDM [8] employs categorical distributions to maintain clear label boundaries, resulting in more precise segmentation outcomes.

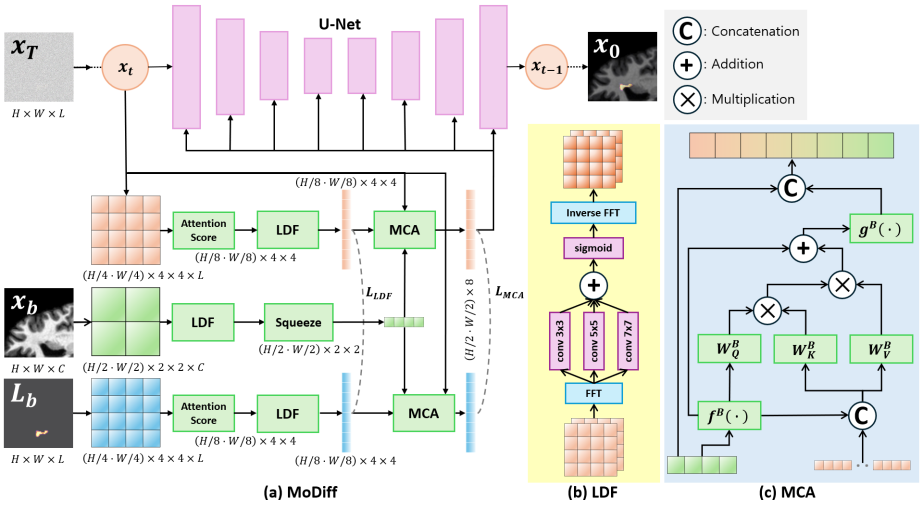

Fig. 1. a) This is a graphical model representing the overall reverse noise process in the training phase of MoDiff. During training, at each step t, the input image b is added to the noisy segmentation label x_t and used as the input image. Using the original label x'_t, b) LDF performs morphological structure extraction and high-frequency noise filtering, and c) the MCA, an attention-based network generating the final conditional image, calculates the loss.

3 Method

3.1 Morphology based Cross Attention Network

The image $x_b \in \mathbb{R}^{H\times W\times C}$ is paired with segmentation labels $l_{b,i}$ (with i denoting multiple independent labels). Instead of randomly selecting a single label for training—which would require i times more iterations—we compute an average probability map L_b by averaging the i labels at each pixel. This improves training efficiency and result stability.

In conditional diffusion models, the condition directly affects performance, yet medical images often exhibit ambiguous boundaries and low contrast (e.g., in MRI or CT), making it difficult to preserve anatomical structure. To address this, we introduce MCA, which enhances morphological feature extraction. Traditional methods (like Canny or Sobel filters) fall short in capturing precise organ contours, and denoising can further degrade label maps.

Our approach learns a high-frequency filter in the frequency domain tailored to each image, enabling effective feature extraction with fewer denoising steps. Specifically, the original image b is divided into 2×2 patches, while the sampled label x_t is subdivided into 4×4 patches. This subdivision reduces high-frequency noise variance, as each patch in b maps to four patches in x_t through linear transformation and positional encoding. The linear mappings used in MCA are

defined as follows:

$$Q = BW_Q, \quad K = SW_K, \quad V = SW_V, \tag{1}$$

where$B = \{x_b^1, x_b^2, \ldots, x_b^n\}$ is the set of patches of the original image, and$S$ is the set of patches of the sampled label with LDF applied. W_Q, W_K, and W_V are learnable projection matrices. Attention scores calculated for each small patch are concatenated to form $S = [x_t^1, x_t^2, \ldots, x_t^m]$. The queries and keys are linearly projected vectors, $Q \in \mathbb{R}^{H/2 \cdot W/2 \cdot 4}$, $K \in \mathbb{R}^{H/4 \cdot W/4 \cdot 16}$, and $V \in \mathbb{R}^{H/4 \cdot W/4 \cdot 16}$. Sinusoidal positional encoding is also added to represent the fixed position of each patch. The attention score $S' = [x'^1_t, x'^2_t, \ldots, x'^m_t]$ for each key and value pair is calculated as follows:

$$x'^m_t = \frac{(x_t^m W_q)(x_t^m W_k)^\top}{\sqrt{d}}, \tag{2}$$

where x'^m_t is the attention score embedded for the m-th label patch, W_q and W_k are learnable weight matrices, and d is the dimension of the key vector K. Softmax normalization is omitted to maximize the contrast between noise values and actual label values, enhancing noise removal performance. The MCA encoder E, utilizing the computed S and B, operates as follows:

$$y^B = [f(B) \,\|\, S'], \tag{3}$$

$$O^B = f^B(\text{Squeeze}(\text{LDF}(B))) + \text{MSA}\big(\text{LN}\big(\text{LDF}(y^B)\big)\big), \tag{4}$$

$$z^B = [g^B(O^B) \,\|\, B], \tag{5}$$

where f^B is an embedding function, g^B is a mapping function, and $\|$ denotes vector concatenation. LN represents layer normalization, and MSA is multi-head self-attention. MSA additionally applies softmax normalization to the formula for calculating the attention score of S, and the denominator includes multiplied by the number of heads h. In this study, the number of heads used is 4. The graphical model of proposed approach is illustrated in Figure 1.

3.2 Learnable Discrete Frequency Filter

To further enhance morphological feature extraction and effectively remove noise, we introduce the LDF into MCA. LDF transforms MCAs output into the frequency domain, applies convolution and a sigmoid function to create a high-frequency filter, and then transforms it back to the spatial domain for noise removal. The LDF process is defined as follows:

$$\text{LDF}(y^B) = \mathcal{F}^{-1}\left(\sigma\left(\mathcal{F}(y^B) * H\right)\right), \tag{6}$$

where y^B is the input or intermediate output of the MCA encoder, representing the embedding vectors of the original image and sampled label patches. $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the FFT and its inverse, respectively. H is a learnable high-frequency

filter, which undergoes convolution and is normalized with a sigmoid function σ. The convolution is applied hierarchically with filter sizes of 3x3, 5x5, and 7x7, and a stride value is used to preserve the shape of the input image. The calculated attention score is used as a condition for each reverse diffusion process and is concatenated with x_t for training. To prevent overfitting, the original image is also concatenated to the generated c_t, which is then used to infer the label for the next step.

3.3 Training

The diffusion model's primary loss, $\mathcal{L}_{\text{simple}}$, is defined as the mean squared error between the predicted noise ϵ_θ and the actual noise ϵ. However, this loss alone does not suffice for low-contrast, single-channel medical images. Therefore, auxiliary loss functions based on LDF and MCA are introduced.

For LDF, to ensure the generated feature map preserves the original image's morphological details, we add an auxiliary loss:

$$\mathcal{L}_{\text{LDF}} = \mathbb{E}\left[\left\|\text{LDF}(y^B) - y^B_{\text{target}}\right\|^2\right], \tag{7}$$

where y^B_{target} is the ideal morphological feature map. This term drives LDF to produce feature maps that closely match the desired structural details.

For MCA, an attention regularization loss, $\mathcal{L}_{\text{MCA}}$, is introduced to control the distribution of attention scores. This helps prevent bias toward specific patches and ensures the attention mechanism effectively learns the relationship between the original image and its sampled label.

The overall loss function is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} + \lambda_1 \mathcal{L}_{\text{LDF}} + \lambda_2 \mathcal{L}_{\text{MCA}}, \tag{8}$$

with $\lambda_1 = \lambda_2 = 0.5$. This configuration improves noise removal while enhancing morphological feature extraction and balanced attention learning.

During training, the model parameters are updated via backpropagation to minimize $\mathcal{L}_{\text{total}}$, thereby simultaneously boosting denoising capability and segmentation performance.

4 Experiments

4.1 Dataset and Experimental setup

Lung Image Database Consortium image collection (LIDC-IDRI) We used the LIDC-IDRI dataset [14], comprising 1,018 low-dose chest CT scans from 1,010 subjects. Lung nodules were annotated by 12 radiologists, with each nodule labeled by four radiologists based on location, size, shape, and malignancy. The dataset provides 128×128 pixel 2D patches centered on each nodule. The training set includes 14,000 lesion images, while the test set has 1,096 images.

Multiple Sclerosis Lesion Segmentation (MS-MRI) This dataset [15] contains longitudinal MRI scans from five subjects focused on white matter lesions in multiple sclerosis. Two radiologists independently annotated the lesions, and each slice includes four modalities: proton density, FLAIR, MP-RAGE, and T2-weighted images. It is split into 5,661 training and 767 test lesion images, with preprocessed 128×128 pixel patches centered on the lesions.

Implementation Details The proposed method was implemented using the PyTorch framework and a 4-way RTX 3090 setup. For CCDM and our model, we set the time step $T = 250$, while for other diffusion models, the time step was set to $T = 1000$ with a linear noise schedule. The optimizer used was Adam, with a learning rate of $1e^{-4}$ for MoDiff. Additionally, the scheduler was set to ReduceLROnPlateau with a factor of 0.2 and patience of 10. For the comparison models, we used the parameters provided in their respective published code or papers. For all models, the binary masks for each image were generated by thresholding the predicted probability maps at 0.9, retaining only the pixels with a probability of 0.9 or higher.

Evaluation Metrics We evaluate performance using several metrics: GED [17–19] and HM-IoU [20, 21] assess distribution differences between generated and ground-truth label maps, NCC [22] measures image similarity, CI [2] synthesizes various evaluation criteria. Each metric is computed over n samples (e.g., GED_n), with larger n yielding more precise estimates.

4.2 Comparison with the Baseline Methods

Quantitative Comparison As shown in Table 1, 2, Our model consistently achieves lower GED_n and higher HM-IoU_n scores compared to other models, indicating improved performance with lower GED_n and higher HM-IoU_n values. Additionally, MoDiff demonstrates competitive performance in terms of NCC_n and CI_n scores against state-of-the-art models, where higher NCC_n and CI_n values reflect better performance. Overall, these results demonstrate MoDiff not only aligns more closely with the ground truth label distribution but also maintains high structural similarity and diversity in the generated samples, outperforming baseline methods across multiple evaluation metrics. Furthermore, performance improves for all models when using 32 sampled labels with different noise levels compared to 16 sampled labels. This suggests the model's diversity and ability to generate labels over a wider area contribute to enhanced performance.

Qualitative Comparison Figure 2 illustrates visual comparisons between the proposed MoDiff model and other baseline methods. As depicted, MoDiff captures finer details and exhibits smoother boundaries, particularly when dealing

Table 1. Quantitative comparison of results with state-of-the-art ambiguous segmentation networks in terms of GED, HM-IoU, NCC and CI on the LIDC-IDRI dataset. The best results are shown in bold, achieving state-of-the-art performance across all evaluation scores. Additionally, cases with a higher number of samples show relatively better performance.

Method	GED_{16}	$HM\text{-}IoU_{16}$	NCC_{16}	CI_{16}	GED_{32}	$HM\text{-}IoU_{32}$	NCC_{32}	CI_{32}
Prob.Unet [16]	0.3050	0.5572	0.4795	0.7312	0.2998	0.5501	0.4810	0.7366
cFlow [6]	0.2252	0.5802	0.5234	0.6846	0.2258	0.5857	0.5289	0.6952
MoSE [7]	0.2176	0.6202	0.3900	0.7258	0.2096	0.6225	0.3987	0.7280
CIMD [2]	0.2198	0.6125	0.4541	0.7482	0.2109	0.6184	0.4602	0.7543
CCDM [8]	0.1690	0.7922	0.5553	0.8362	0.1678	0.8107	0.5570	0.8590
MoDiff (*ours*)	**0.1159**	**0.8328**	**0.5929**	**0.8892**	**0.1163**	**0.8588**	**0.5953**	**0.8912**

Table 2. Quantitative comparison of results with state-of-the-art ambiguous segmentation networks in terms of GED, HM-IoU, NCC and CI on the MS-MRI dataset. The best results are shown in bold, achieving state-of-the-art performance across all evaluation scores. Additionally, cases with a higher number of samples show relatively better performance.

Method	GED_{16}	$HM\text{-}IoU_{16}$	NCC_{16}	CI_{16}	GED_{32}	$HM\text{-}IoU_{32}$	NCC_{32}	CI_{32}
Prob.Unet [16]	0.2956	0.7430	0.3734	0.6635	0.2573	0.7464	0.3778	0.6597
cFlow [6]	0.2674	0.7513	0.3341	0.7148	0.2458	0.7611	0.3411	0.7322
MoSE [7]	0.3278	0.7054	0.3858	0.6223	0.3594	0.7154	0.3870	0.6332
CIMD [2]	0.3984	0.6864	0.3247	0.6623	0.3566	0.6401	0.3276	0.6720
CCDM [8]	0.2306	0.8405	0.4005	0.7443	0.2289	0.8488	0.4013	0.7527
MoDiff (*ours*)	**0.1671**	**0.8485**	**0.4109**	**0.7932**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

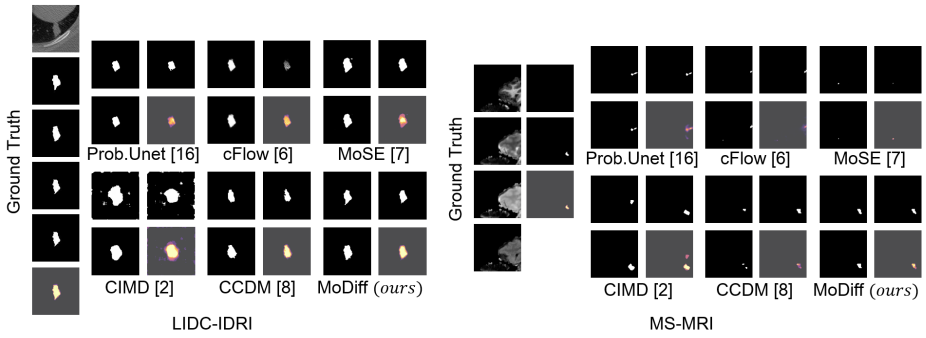

Fig. 2. The lesion segmentation results of the selected MoDiff and previous models for experiments are visualized. The left side shows the ground truth and the average label.

with complex structures, outperforming other models in terms of visual fidelity. This qualitative improvement complements the quantitative results and underscores the effectiveness of MoDiff in generating high-quality samples closely resembling ground truth images.

Table 3. Ablation Study results of hyperparameters on all the datasets.

Module		LIDC-IDRI				MS-MRI			
MCA	LDF	GED	HM-IoU	NCC	CI	GED	HM-IoU	NCC	CI
✓		0.1432	0.7954	0.4933	0.8281	0.2335	0.7649	0.3394	0.7428
	✓	0.1339	0.8191	0.5231	0.8377	0.2116	0.7748	0.3561	0.7502
✓	✓	**0.1163**	**0.8588**	**0.5953**	**0.8912**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

Table 4. Case Study results of Edge Detection Methods on all the datasets.

Method	LIDC-IDRI				MS-MRI			
	GED	HM-IoU	NCC	CI	GED	HM-IoU	NCC	CI
Canny	0.2544	0.7588	0.4855	0.7590	0.2929	0.5883	0.3552	0.6313
Gaussian	0.2465	0.7642	0.5444	0.7469	0.2814	0.6102	0.3893	0.6954
Roberts	0.2208	0.8034	0.4902	0.6820	0.2738	0.6659	0.3617	0.7512
Prewitt	0.2265	0.8061	0.5047	0.7469	0.2549	0.7091	0.3859	0.7427
Sobel	0.2044	0.8222	0.5331	0.7590	0.2126	0.7218	0.3996	0.7843
LDF (*ours*)	**0.1163**	**0.8588**	**0.5953**	**0.8912**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

4.3 Ablation and Case Study

Tables 3 present the ablation study results on the LIDC-IDRI and MS-MRI datasets, assessing the individual and combined contributions of the MCA and LDF modules. When using MCA alone (first row), the performance is the lowest, indicating its limited ability to capture detailed structural information. Employing only LDF (second row) improves several metrics, such as HM-IoU and NCC, yet does not reach optimal performance. Notably, the joint use of MCA and LDF (third row) yields the best results across all evaluation metrics, demonstrating that these modules complement each other to enhance both segmentation accuracy and consistency.

Tables 4 provide a comparative case study of various edge detection methods on the same datasets, including traditional approaches and the proposed LDF method. The results clearly show that while conventional edge detectors achieve moderate performance, they tend to concentrate weights on boundary pixels around organ contours or other edge regions, thereby failing to fully utilize the information from adjacent pixels. These findings underscore the superiority of the

LDF method in accurately capturing edge details and maintaining segmentation consistency on both LIDC-IDRI and MS-MRI datasets.

5 Conclusion

MoDiff is a morphology-emphasized diffusion model that tackles the challenges of ambiguous medical image segmentation by integrating probability-based label maps with a learnable morphological frequency space. This design enables the model to capture inherent uncertainties and fine structural details across multiple annotations, resulting in consistent and precise segmentation. Extensive evaluations on the LIDC-IDRI and MS-MRI datasets demonstrate that MoDiff outperforms conventional probabilistic models in segmentation accuracy, boundary delineation, and sample consistency. While increased sampling steps have led to improved performance metrics, this comes at the expense of computational efficiency. To address this limitation, future work will focus on refining the model architecture—exploring approaches such as noise scheduling and multimodal structures to reduce the number of required sampling steps.

Acknowledgments. This work was supported by an IITP grant funded by the Korean government MSIT (No. RS-2025-02304331, Digital Columbus Project).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015)
2. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous Medical Image Segmentation Using Diffusion Models. In: CVPR, pp. 11536–11546 (2023)
3. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems **33**, 12756–12767 (2020)
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015, pp. 234–241. Springer (2015)
5. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
6. Selvan, R., Faye, F., Middleton, J., Pai, A.: Uncertainty quantification in medical image segmentation with normalizing flows. In: MLMI 2020 (held with MICCAI 2020), pp. 80–90. Springer (2020)
7. Gao, Z., Chen, Y., Zhang, C., He, X.: Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. In: ICLR (2023).
8. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: ICCV, pp. 1119–1129 (2023)

9. Chen, T., Wang, C., Shan, H.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: MICCAI, pp. 491–501. Springer (2023)
10. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning, pp. 1623–1639. PMLR (2024)
11. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)
13. Zhang, S., Wang, S., Miao, H., Chen, H., Fan, C., Zhang, J.: Score-CDM: Score-weighted convolutional diffusion model for multivariate time series imputation. arXiv preprint arXiv:2405.13075 (2024)
14. Armato III, S.G., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Yankelevitz, D., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., et al.: Lung image database consortium: Developing a resource for the medical imaging research community. Radiology **232**(3), 739–748 (2004)
15. Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., et al.: Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage **148**, 77–102 (2017)
16. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M., Jimenez Rezende, D., Ronneberger, O.: A probabilistic U-net for segmentation of ambiguous images. Advances in Neural Information Processing Systems **31** (2018)
17. Bellemare, M.G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., Munos, R.: The Cramer distance as a solution to biased Wasserstein gradients. arXiv preprint arXiv:1705.10743 (2017)
18. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving GANs using optimal transport. arXiv preprint arXiv:1803.05573 (2018)
19. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. J. Stat. Plan. Inference **143**(8), 1249–1272 (2013).
20. Kassapis, E., Dikov, G., Gupta, D.K., Nugteren, C.: Calibrated adversarial refinement for stochastic semantic segmentation. In: ICCV, pp. 7057–7067 (2021)
21. Kohl, S.A.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S.M., Kohli, P., Zisserman, A., Ronneberger, O.: A hierarchical probabilistic U-net for modeling multi-scale ambiguities. arXiv preprint arXiv:1905.13077 (2019)
22. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: PHiSeg: Capturing Uncertainty in Medical Image Segmentation. In: MICCAI 2019, pp. 119–127. Springer (2019)