Modality Translation and Registration of MR and Ultrasound Images Using Diffusion Models

Xudong Ma[1], Nantheera Anantrasirichai[1], Stefanos Bolomytis[2], Alin Achim[1]

[1]Visual Information Laboratory, University of Bristol, Bristol, the United Kingdom

[2]Southmead Hospital, North Bristol NHS Trust, Bristol, the United Kingdom

{xudong.ma, n.anantrasirichai, alin.achim}@bristol.ac.uk, stefanos.bolomytis@nbt.nhs.uk

Abstract—Multimodal MR-US registration is critical for prostate cancer diagnosis. However, this task remains challenging due to significant modality discrepancies. Existing methods often fail to align critical boundaries while being overly sensitive to irrelevant details. To address this, we propose an anatomically coherent modality translation (ACMT) network based on a hierarchical feature disentanglement design. We leverage shallow-layer features for texture consistency and deep-layer features for boundary preservation. Unlike conventional modality translation methods that convert one modality into another, our ACMT introduces the customized design of an intermediate pseudo modality. Both MR and US images are translated toward this intermediate domain, effectively addressing the bottlenecks faced by traditional translation methods in the downstream registration task. Experiments demonstrate that our method mitigates modality-specific discrepancies while preserving crucial anatomical boundaries for accurate registration. Quantitative evaluations show superior modality similarity compared to state-of-the-art modality translation methods. Furthermore, downstream registration experiments confirm that our translated images achieve the best alignment performance, highlighting the robustness of our framework for multi-modal prostate image registration.

Index Terms—Modality Translation, MR-US Registration, Prostate Cancer, Diffusion Models, Multimodal Image Registration, Unsupervised Learning

I. INTRODUCTION

Prostate cancer is a leading cause of death among men worldwide [1], with accurate diagnosis relying heavily on advanced imaging techniques. Magnetic Resonance Imaging (MRI) and Ultrasound (US) are widely used in prostate cancer imaging, each offering unique advantages: MRI provides high soft-tissue contrast for lesion detection, while US enables real-time guidance for biopsies. However, their significant differences in anatomical representation pose a major challenge for multi-modal registration, which is essential for combining their complementary strengths.

To address modality discrepancies, some methods employ image segmentation, aligning segmented regions to avoid direct cross-modal registration challenges [2], [3]. However, these approaches require extensive annotated data, which are labor-intensive and scarce, limiting their practical application. To overcome these limitations, we previously proposed a Partial Modality Translation (PMT) approach [6], which established an intermediate modality to bridge MRI and US domains. While PMT improved texture similarity, it does not truly achieve customized modality translation and retains excessive details irrelevant to registration. This oversight left room for improvement in the registration results.

Building on this foundation, we propose a novel hierarchical feature disentanglement method for anatomically coherent modality translation (ACMT). Through a customized intermediate modality design, ACMT logically fulfills the goal that PMT failed to achieve. The ACMT advances PMT by addressing two key limitations: (1) explicitly preserving boundary information while maintaining texture consistency, and (2) relaxing the requirement for photorealism so as to reduce the preservation of unnecessary and overly detailed textures within the prostate. Specifically, we leverage the distinct characteristics of shallow and deep features in convolutional networks. Shallow-layer features, which primarily capture low-level texture information, are processed with larger convolutional kernels. This design is motivated by the fact that larger kernels can effectively model broader texture patterns, which are essential for achieving cross-modal consistency [14]. In contrast, deep-layer features, which are more sensitive to high-level structural details such as shapes, are processed with smaller convolutional kernels. Smaller kernels are better suited for capturing fine-grained details and precise anatomical structures, making them ideal for boundary preservation [8]. Additionally, we enhance boundary extraction using Sobel filtering to further improve structural alignment [10]. In addition, by removing the adversarial loss used in PMT, we further reduce the emphasis on photorealism, as the intermediate modality does not require highly realistic details. This design ensures that the output retains only the boundary information most relevant for registration, while promoting texture consistency and suppressing modality-specific anatomical detail discrepancies. As a result, the difficulty of cross-modal registration is significantly reduced.

In the subsequent sections of this paper, we will provide a comprehensive explanation of the mathematical foundations and network architecture of the proposed method in Section 2. Section 3 will present both objective and subjective analyses of the results, demonstrating the effectiveness of our approach. Finally, we will conclude with a summary of our contributions in the last section.

II. PROPOSED METHOD

In this section, we first introduce the theoretical foundation of our method, which is based on a Diffusion Schrödinger

Bridge. Following this, we provide a detailed explanation of the overall workflow of the proposed ACMT framework, as illustrated in Fig. 1.

A. Schrödinger Bridge based Diffusion Model

We design this network by leveraging a Schrödinger Bridge based (SB) diffusion framework [4]. It is inspired by optimal transport (OT) theory [5], [12]. Formally, the SB problem seeks to find the optimal stochastic process that transforms a source distribution P_0 to a target distribution P_1 through intermediate distributions P_t over time t, as defined by:

$$P^{SB} = \{\underset{P_t}{\arg\min}\, D_{KL}(P_t \| W^\sigma)\} \quad with \quad t \sim [0,1], \quad (1)$$

where W^σ denotes the Wiener measure with variance σ. This formulation aims to minimize the Kullback-Leibler (KL) divergence between the process distribution P_t and the reference measure W^σ at each timestep. The collection of optimal distributions $\{P_t^{SB}\}$ constitutes the Schrödinger Bridge connecting P_0^{SB} and P_1^{SB}. Although SB is theoretically a continuous process, Conditional Flow Matching (CFM) formulation allows us to address it in a discretized manner [4], [11]. The CFM formulation establishes that for any two distributions $P_{t_m}^{SB}$ and $P_{t_n}^{SB}$ in the SB where $[t_m, t_n] \subseteq [0,1]$, the intermediate distribution at time $t \in [t_m, t_n]$ follows a Gaussian distribution:

$$p(X_t | X_{t_m}, X_{t_n}) = \mathcal{N}\Big(X_t \Big| w_t X_{t_n} + (1 - w_t) X_{t_m},$$
$$w_t(1 - w_t)\sigma(t_n - t_m)\boldsymbol{I}\Big), \quad (2)$$

where $X_t \sim P_t^{SB}$, $X_{t_m} \sim P_{t_m}^{SB}$, $X_{t_n} \sim P_{t_n}^{SB}$ and $w_t = \frac{t - t_m}{t_n - t_m}$. Furthermore, the joint distribution $P_{t_m t_n}^{SB}$ between any two timesteps can be obtained through an entropy-regularized OT problem:

$$P_{t_m t_n}^{SB} = \underset{P_{t_m, t_n}}{\arg\min}\, E_{(X_{t_m}, X_{t_n})}[\|X_{t_m} - X_{t_n}\|^2]$$
$$- 2\sigma(t_n - t_m)H(X_{t_m}, X_{t_n}), \quad (3)$$

where H denotes the entropy function. This formulation enables the determination of the optimal terminal distribution $P_{t_n}^{SB}$ given any initial distribution $P_{t_m}^{SB}$. Subsequently, using Equation (2), we can compute any intermediate state P_t^{SB} along the Schrödinger Bridge between the $P_{t_m}^{SB}$ and $P_{t_n}^{SB}$.

B. Diffusion Process via CFM formulation

Since the training of our model relies on inputs generated via a diffusion process, we first provide a comprehensive explanation of the diffusion process that forms the foundation of our framework. This diffusion process applies of our modality translation network. However, it is used in evaluation mode with all parameters frozen. To be more specific, we hypothesize the existence of an ideal intermediate modality that focuses primarily on boundary information while disregarding texture details. Through careful network design and loss function formulation, our framework learns to transform both MR and US images into this target modality. This process is achieved by constructing Schrödinger Bridge from either MR P_0^{MR} distribution or US distribution P_0^{US} to a single intermediate modality P_1.

As illustrated in the Diffusion Process (purple block) in Fig. 1, let $x_{t_j} \in P_{t_j}$ represent an intermediate state along the Schrödinger Bridge at time $t_j \in [0,1]$. Our network learns to map x_{t_j} to the terminal state $x_1 \in P_1$. By using the CFM formulation as shown in Equation 2, the next state $x_{t_{j+1}}$ could be computed as:

$$x_{t_{j+1}} = w_{t_{j+1}} x_1 + (1 - w_{t_{j+1}}) x_{t_j} + \mathcal{N}(0, \alpha_{j+1} I), \quad (4)$$

where $w_{t_{j+1}} = \frac{t_{j+1} - t_j}{1 - t_j}$ is the interpolation weight balancing x_1 and x_{t_j}, $\alpha_{j+1} = w_{t_{j+1}}(1 - w_{t_{j+1}})(1 - t_j)\sigma$ controls the noise magnitude, and $\mathcal{N}(0, \alpha_{j+1} I)$ adds Gaussian noise scaled by α_{j+1}.

This iterative process begins with $x_{t_0} = x_0$, the source MR or US image, and progressively transforms it towards x_{t_i}.

C. Training Process

The training process begins with inputs generated by a diffusion process. As shown in the Training Process (left part of Fig.1), during training, we implement the following procedure for each MR-US image pair:

1) Randomly select t_i from the predefined time step pool $\{t_0, t_1, t_2, \ldots, t_T\}$, where each $t_i \in [0,1]$.
2) Using the network in evaluation mode (with all parameters frozen): generate x_{t_i} through the diffusion process as described in the purple block of Fig. 1
3) Switch the network to training mode:
   - Compute the transformation from x_{t_i} to x_1
   - Calculate loss
   - Update network parameters through backpropagation

Through this training strategy, the network learns to transform any intermediate state x_{t_i} to the target modality x_1, effectively capturing the mapping along the entire Schrödinger Bridge. By breaking down the transformation into iterative steps, it ensures gradual refinement of the target modality while preserving critical anatomical information. This approach is particularly advantageous for medical imaging, where the complex relationship between source and target modalities requires robust solutions.

D. Loss Functions

Recent studies have demonstrated the remarkable capability of diffusion models in extracting discriminative features from images [7], [13]. Building upon this observation and the well-established understanding that shallow layers of neural networks are more sensitive to texture information while deeper layers capture boundary details [8], we design a hierarchical feature disentanglement loss to achieve both texture consistency and boundary preservation in our modality translation framework.

As shown in the Hierarchical Feature Disentanglement part of Fig. 1, let f_θ^s and f_θ^d denote the shallow and deep feature
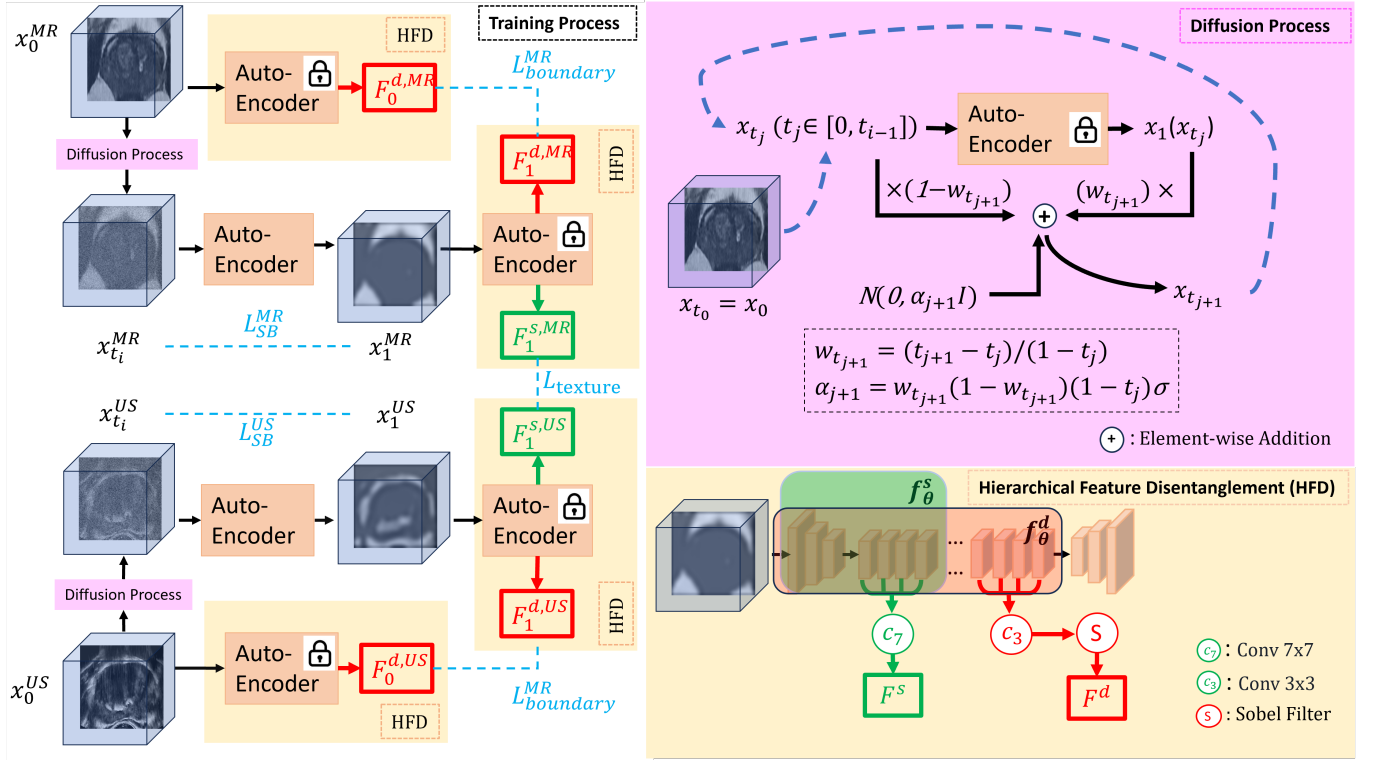
Fig. 1: Hierarchical Feature Disentanglement framework based on a Diffusion Model

extraction functions of our network, respectively. Given an input image x_0 (either MR or US), the shallow and deep features are extracted as:

$$\mathbf{F}_0^s = f_\theta^s(x_0), \quad \mathbf{F}_0^d = f_\theta^d(x_0). \quad (5)$$

Similarly, for the translated image x_1 in the intermediate modality P_1, the corresponding features are:

$$\mathbf{F}_1^s = f_\theta^s(x_1), \quad \mathbf{F}_1^d = f_\theta^d(x_1). \quad (6)$$

1) Texture Consistency Loss: To ensure texture similarity between the translated MR and US images in the intermediate domain, we further process the shallow features $\mathbf{F}_1^{s,MR}$ and $\mathbf{F}_1^{s,US}$ using a large convolutional kernel (7×7) because larger kernels are known to be more effective at capturing general, global features such as texture patterns due to their wider receptive fields. The texture consistency loss is formulated as:

$$\mathcal{L}_{\text{texture}} = \left\| \mathcal{C}_{7\times7}(\mathbf{F}_1^{s,MR}) - \mathcal{C}_{7\times7}(\mathbf{F}_1^{s,US}) \right\|_2^2, \quad (7)$$

where $\mathcal{C}_{7\times7}(\cdot)$ represents the 7×7 convolution operation, and $\|\cdot\|_2^2$ denotes the L2 norm.

2) Boundary Preservation Loss: To preserve the anatomical boundaries of the original images, we process the deep features $\mathbf{F}_0^d$ and $\mathbf{F}_1^d$ using a smaller convolutional kernel (3×3) followed by a Sobel filter. This design choice is motivated by the fact that smaller kernels are particularly effective at extracting fine-grained features, such as edges and boundaries, due to their ability to focus on localized regions and capture

high-frequency details. The boundary preservation loss for MR and US images is defined as:

$$\mathcal{L}_{\text{boundary}}^{MR} = \left\| \mathcal{S}(\mathcal{C}_{3\times3}(\mathbf{F}_1^{d,MR})) - \mathcal{S}(\mathcal{C}_{3\times3}(\mathbf{F}_0^{d,MR})) \right\|_1, \quad (8)$$

$$\mathcal{L}_{\text{boundary}}^{US} = \left\| \mathcal{S}(\mathcal{C}_{3\times3}(\mathbf{F}_1^{d,US})) - \mathcal{S}(\mathcal{C}_{3\times3}(\mathbf{F}_0^{d,US})) \right\|_1, \quad (9)$$

where $\mathcal{C}_{3\times3}(\cdot)$ denotes the 3×3 convolution operation, $\mathcal{S}(\cdot)$ represents the Sobel filter, and $\|\cdot\|_1$ is the L1 norm. The total boundary preservation loss is:

$$\mathcal{L}_{\text{boundary}} = \frac{1}{2}\left(\mathcal{L}_{\text{boundary}}^{MR} + \mathcal{L}_{\text{boundary}}^{US}\right). \quad (10)$$

3) SB Loss: To ensure that the translation process follows the optimal transport path defined by the Schrödinger Bridge, we introduce the SB constraint loss based on the joint distribution $P_{t_i,1}^{SB}$ following the theory in equation 3. For MR and US images, the SB loss is defined as:

$$\mathcal{L}_{\text{SB}}^{MR}(\theta_i, t_i) = \mathbb{E}_{(x_{t_i}^{MR}, x_1^{MR})}\left[\|x_{t_i}^{MR} - x_1^{MR}\|^2\right] \\ - 2\sigma(1-t_i)H(x_{t_i}^{MR}, x_1^{MR}), \quad (11)$$

$$\mathcal{L}_{\text{SB}}^{US}(\theta_i, t_i) = \mathbb{E}_{(x_{t_i}^{US}, x_1^{US})}\left[\|x_{t_i}^{US} - x_1^{US}\|^2\right] \\ - 2\sigma(1-t_i)H(x_{t_i}^{US}, x_1^{US}). \quad (12)$$

where $x_1^{MR} = f_{\theta_i}(x_1^{MR}|x_{t_i}^{MR})$ and $x_1^{US} = f_{\theta_i}(x_1^{US}|x_{t_i}^{US})$ are the terminal states predicted by the network. The total SB loss is then:

$$\mathcal{L}_{\text{SB}} = \frac{1}{2}\left(\mathcal{L}_{\text{SB}}^{MR} + \mathcal{L}_{\text{SB}}^{US}\right). \quad (13)$$

The overall loss function combines the texture consistency loss, boundary preservation loss, and the SB loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{texture}}\mathcal{L}_{\text{texture}} + \lambda_{\text{boundary}}\mathcal{L}_{\text{boundary}} + \lambda_{\text{SB}}\mathcal{L}_{\text{SB}} \quad (14)$$

The weighting coefficients λ_{texture}, $\lambda_{\text{boundary}}$, and λ_{SB} are carefully tuned to balance the contributions of each loss component, ensuring that the network simultaneously achieves texture consistency, boundary preservation, and optimal transport.

III. RESULTS AND EVALUATION

1) Dataset: 3D MR and US paired images were provided by Southmead Hospital Bristol. The MR images are T2-weighted volumes, while the 3D US images are derived from biopsy procedures. To ensure consistency, all volumes were preprocessed to focus on the prostate region while retaining relevant surrounding structures. Additionally, to enhance generalizability and robustness, we applied data augmentation techniques such as flipping and rotation. To balance computational feasibility and anatomical fidelity, all volumes were standardized to a resolution of $128 \times 128 \times 64$ pixels. The dataset was partitioned into an 80% training set and a 20% test set, and all experiments employed cross-validation.

2) Evaluation: We evaluated both the quality of modality translation and the corresponding registration performance using the same registration model, FSDiffReg [9]. We compared: (1) registration on original MR-US images, and (2) registration on modality-translated results from state-of-the-art (SOTA) approaches, including UNSB [4], an ICLR 2024 modality translation method, and PMT [6], our previously proposed method that focuses on texture similarity.

To assess the quality of modality translation, we employed two widely-used metrics: FID and KID. For registration evaluation, our experts manually segmented the prostate on several key frames for each test case to create ground-truth masks. The generated deformation fields, based on different modality translation methods, were applied to warp these masks, and the results were scored with:

- **Dice Similarity Coefficient (DSC)**: Defined as $\frac{2|X \cap Y|}{|X| + |Y|}$, this metric quantifies the volume overlap between the warped mask X and the ground-truth mask Y.
- **Intersection-over-Union (IoU)**: Given by $\frac{|X \cap Y|}{|X \cup Y|}$, IoU provides a stricter assessment of boundary alignment by measuring the ratio of the shared to the combined volume.
- **Average Surface Distance (ASD)**: This metric reports the mean Euclidean distance between the surfaces of the warped and ground-truth segmentations.

The experimental results demonstrate that our method achieves the best FID and KID scores, as shown in Table I. Specifically, our approach reduces the FID by 65.91% and the KID by 83.93%, outperforming UNSB approximately 10 times and 12 times, respectively. Compared to our previously proposed PMT method, our current method shows further improvements in both metrics, highlighting its superior performance in modality translation. On the other hand, as shown in Table II, our ACMT framework attains the best overall

TABLE I: Quantitative evaluation of modality translation quality using FID and KID (lower is better).

Method	FID ↓ (decrease by ↑)	KID ↓ (decrease by ↑)
Original	404.88	0.56
UNSB	377.92 (6.66%)	0.52 (7.14%)
PMT	170.02 (58.01%)	0.11 (80.36%)
ACMT(Ours)	**138.01 (65.91%)**	**0.09 (83.93%)**

TABLE II: Quantitative evaluation of registration performance based on different modality translation methods, all registered using a consistent approach.

Method	DSC ↑	IoU ↑	ASD ↓
Original	0.92	0.87	10.74
UNSB	0.92	0.88	12.83
PMT	**0.95**	**0.91**	9.18
ACMT (ours)	**0.95**	0.90	**6.82**

registration results, leading all methods in DSC and ASD. While its IoU is only 0.01 below that of PMT, ACMT lowers ASD by 25.7 % compared with the runner-up, demonstrating superior surface alignment. These findings confirm that ACMT produces more anatomically faithful intermediate representations and yields substantially higher registration accuracy than either UNSB or PMT-based variants.

Visually, we randomly selected two patients to illustrate the modality translation and image registration comparison. As shown in Fig. 2, Our method maximizes the similarity within the prostate interior while preserving boundary information that is most relevant for registration. At the same time, it suppresses unnecessary internal texture details that could mislead the registration process. This advantage is further validated in the registration results shown in Fig. 3. In particular, in the second row, within the red boxes, our method is the only one that achieves relatively smooth registration, while all other methods exhibit noticeable discontinuities. Similarly, in the first patient case (top row), although all modality translation approaches help improve the smoothness in the red-boxed region, the yellow boxes highlight that our method introduces the least over-deformation. This suggests that our estimated deformation field is the most realistic and anatomically plausible among all methods.

IV. CONCLUSION

In this work, we propose a novel ACMT method for modality translation between MR and US images using a hierarchical feature disentanglement idea. We leverage shallow features to ensure texture consistency and deep features to preserve anatomical boundaries, resulting in anatomically coherent pseudo-representations. Our unsupervised framework achieves customized modality translation, effectively removing irrelevant information from source images that would otherwise hinder cross-modal registration. Experimental results show that our framework consistently outperforms SOAT methods in both modality translation and registration, achieving superior
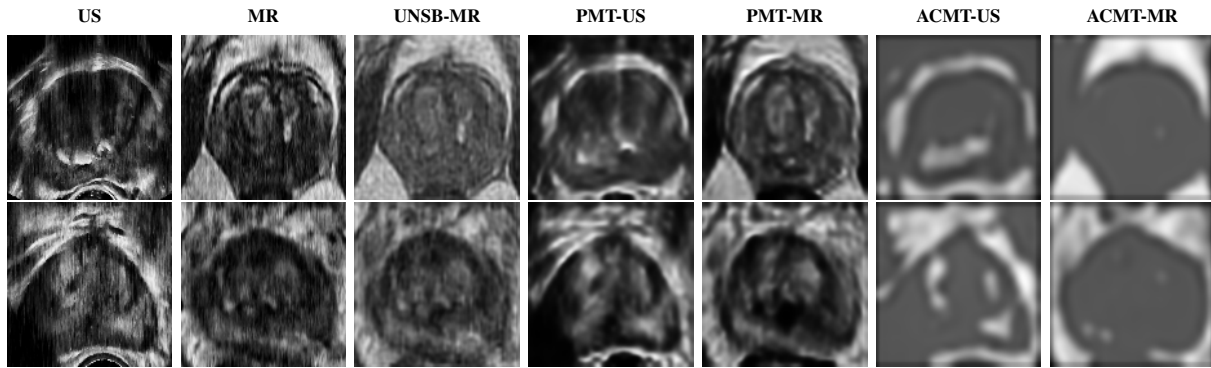
Fig. 2: Modality translation results for two patients (two rows), showing original US and MR images, UNSB translation, and intermediate translation from PMT and ACMT.
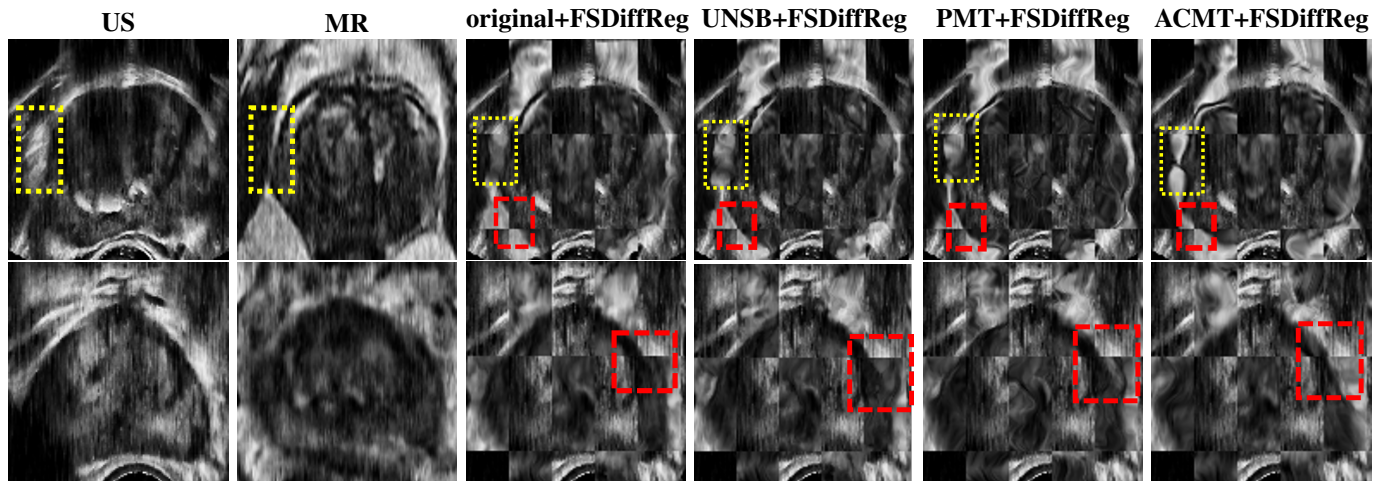

Fig. 3: Registration results for two patients. Each row corresponds to one patient and displays the original US and MR images, followed by registration results using FSDiffReg applied to the original MR-US pair, as well as to images translated by UNSB, PMT, and our ACMT methods. Results are shown in a chessboard layout for visual comparison.

performance in both quantitative metrics and visual assessments.

REFERENCES

[1] Bratt, O., *et al.*: Population-based organised prostate cancer testing: results from the first invitation of 50-year-old men. European Urology **85**(3), 207–214 (2024)

[2] Chen, Y., Xing, L., Yu, L., Liu, W., Pooya Fahimian, B., Niedermayr, T., Bagshaw, H.P., Buyyounouski, M., Han, B.: Mr to ultrasound image registration with segmentation-based learning for hdr prostate brachytherapy. Medical physics **48**(6), 3074–3083 (2021)

[3] Jiang, J., Guo, Y., Bi, Z., Huang, Z., Yu, G., Wang, J.: Segmentation of prostate ultrasound images: the state of the art and the future directions of segmentation algorithms. Artificial Intelligence Review **56**(1), 615–651 (2023)

[4] Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural schrödinger bridge. In: ICLR (2024)

[5] Léonard, C.: A survey of the schrodinger problem and some of its connections with optimal transport. Dynamical Systems **34**(4), 1533–1574 (2014)

[6] Ma, X., Anantrasirichai, N., Bolomytis, S., Achim, A.: Pmt: Partial-modality translation based on diffusion models for prostate magnetic resonance and ultrasound image registration. In: Annual Conference on Medical Image Understanding and Analysis. pp. 285–297. Springer (2024)

[7] Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10619–10629 (2022)

[8] Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H.I.K., Frisky, A.Z.K., ul Haq, Q.M.: Understanding of convolutional neural network (cnn): A review. International Journal of Robotics and Control Systems **2**(4), 739–748 (2022)

[9] Qin, Y., Li, X.: Fsdiffreg: Feature-wise and score-wise diffusion-guided unsupervised deformable image registration for cardiac images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–665. Springer (2023)

[10] Ranjan, R., Avasthi, V.: Edge detection using guided sobel image filtering. Wireless Personal Communications **132**(1), 651–677 (2023)

[11] Tong, A., *et al.*: Conditional flow matching: Simulation-free dynamic optimal transport. arXiv preprint arXiv:2302.00482 **2**(3) (2023)

[12] Wang, G., *et al.*: Deep generative learning via schrödinger bridge. In: International conference on machine learning. pp. 10794–10804. PMLR (2021)

[13] Yang, X., Wang, X.: Diffusion model as representation learner. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18938–18949 (2023)

[14] Youwang, K., Oh, T.H., Pons-Moll, G.: Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4347–4356 (2024)