# S-CycleGAN: Semantic Segmentation Enhanced CT-Ultrasound Image-to-Image Translation for Robotic Ultrasonography

Yuhan Song and Nak Young Chong

***Abstract*— Ultrasound imaging is pivotal in various medical diagnoses due to its non-invasive nature and safety. In clinical practice, the accuracy and precision of ultrasound image analysis are critical. Recent advancements in deep learning are showing great capacity of processing medical images. However, the data hungry nature of deep learning and the shortage of high-quality ultrasound image training data suppress the development of deep learning based ultrasound analysis methods. To address these challenges, we introduce an advanced deep learning model, dubbed S-CycleGAN, which generates high-quality synthetic ultrasound images from computed tomography (CT) data. This model incorporates semantic discriminators within a CycleGAN framework to ensure that critical anatomical details are preserved during the style transfer process. The synthetic images are utilized to enhance various aspects of our development of the robot-assisted ultrasound scanning system. The data and code will be available at https://github.com/yhsong98/ct-us-i2i-translation.**

## I. INTRODUCTION

Ultrasound imaging is one of the most widely implemented medical imaging modalities, offering a versatile, non-invasive, and cost-effective method for visualizing the internal structures of the body in real-time. Although ultrasound imaging is safe and convenient, analyzing these images presents considerable challenges due to factors such as low contrast, acoustic shadows, and speckles [1]. Deep learning based medical image processing methods have made great breakthroughs in recent years and have been the state-of-the-art tool for medical image processing applications in various fields, including detection, segmentation, classification, and synthesis [2].

Nonetheless, due to the data-hungry nature of deep learning, the performance of those methods relies heavily on a large amount of image data and manual annotations. While progress in unsupervised learning techniques and the emergence of large-scale open-source image datasets have mitigated these issues somewhat, these solutions are less applicable in the field of medical image processing due to several factors [3]. Firstly, medical images require precise and reliable annotations, which must often be provided by expert clinicians, making the process time-consuming and expensive. Secondly, patient privacy concerns limit the availability and sharing of medical datasets. Moreover, the variability in medical imaging equipment and protocols across different healthcare facilities can lead to inconsistencies in the data, complicating the development of generalized models.

Along the lines, we are building a fully automated robot-assisted ultrasound scan system (RUSS). This platform is designed to perform abdominal ultrasound scans without any human intervention (Fig. 1). Thus we have proposed several versions of ultrasound image segmentation algorithms as evaluation metrics for the robot arm movements [4], [5], [6]. However, our prior efforts have been restricted by limited data sources. While our segmentation algorithms have demonstrated effectiveness within our experimental datasets, we anticipate that training our model with a more diverse array of data would enhance its robustness and general applicability. Furthermore, we aim to create a simulation environment to facilitate the development of our RUSS, allowing for refined testing and optimization under controlled conditions. A pre-operative 3D model reconstructed from CT scans are planned to be utilized as the scan target. Based on the current contact point and angle of the virtual ultrasound probe, the system will generate and provide a corresponding ultrasound image as feedback. This integration will enable the RUSS to simulate realistic scanning scenarios, allowing for precise alignment and positioning adjustments that reflect actual clinical procedures.

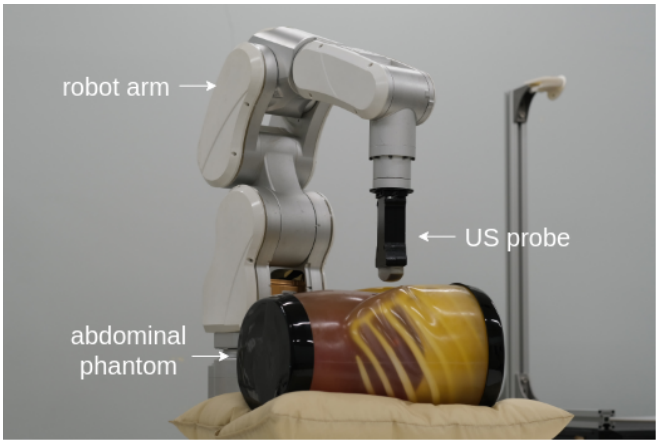



Fig. 1: Our robot-assisted ultrasound imaging system

In this research, we proposed a semantically enhanced CycleGAN, dubbed S-CycleGAN. By adding additional segmentation models as semantic discriminators, together with the original style discriminator, the proposed model is capable of transferring the style of CT slice to the ultrasound domain while keeping the transformed image semantically consistent with the source image.

This work was supported by JSPS KAKENHI Grant Number JP23K03756.

Both authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan {yuhan-s, nakyoung}@jaist.ac.jp

## II. RELATED WORK

### A. Image-to-image translation

The process of image-toimage translation is fundamental in various applications, ranging from artistic style transfer to synthesizing realistic datasets. One seminal work in this field is the introduction of the Generative Adversarial Network (GAN) by Goodfellow et al. [7]. The GAN framework involves a dual-network architecture where a generator network competes against a discriminator network, fostering the generation of highly realistic images. Building on this, Zhu et al. introduced CycleGAN [8], which allows for image-to-image translation in the absence of paired examples. In the context of medical imaging, Sun et al. [9] leveraged a double U-Net CycleGAN to enhance the synthesis of CT images from MRI images. Their model incorporates a U-Net-based discriminator that improves the local and global accuracy of synthesized images. Chen et al. [10] introduced a correction network module based on an encoder-decoder structure into a CycleGAN model. Their module incorporates residual connections to efficiently extract latent feature representations from medical images and optimize them to generate higher-quality images.

### B. Ultrasound image synthesis

As for medical ultrasound image synthesis, advancements have been achieved due to the integration of deep learning techniques, particularly GANs and Denoising Diffusion Probabilistic Models (DDPMs) [11]. Liang et al. [12] employed GANs to generate high-resolution ultrasound images from low-resolution inputs, thereby enhancing image clarity and detail that are crucial for effective medical analysis. Stojanovski et al. [13] introduced a novel approach to generating synthetic ultrasound images through DDPM. Their study leverages cardiac semantic label maps to guide the synthesis process, producing realistic ultrasound images that can serve as substitutes for actual data in training deep learning models for tasks like cardiac segmentation.

In the specific context of synthesizing ultrasound images from CT images, Vitale et al. [14] proposed a two-stage pipeline. Their method begins with the generation of intermediate synthetic ultrasound images from abdominal CT scans using a ray-casting approach. Then a CycleGAN framework operates by training on unpaired sets of synthetic and real ultrasound images. Song et al. [15] also proposed a CycleGAN based method to synthesize ultrasound images from abundant CT data. Their approach leverages the rich annotations of CT images to enhance the segmentation network learning process. The segmentation networks are initially pretrained on the synthetic dataset translated from preprocessed CT images. Then they are fine-tuned on actual ultrasound images to refine their ability to accurately segment kidneys.

## III. METHODOLOGY

### A. CycleGAN

In this research, we aim to translate CT images into ultrasound images. Conventionally, one might consider training a neural network that inputs a CT image and outputs its corresponding ultrasound image, followed by computing the similarity between the synthesized ultrasound image and the actual ultrasound image to update the network. However, we face a challenge: our datasets, one containing abdominal CT volumes [16] and the other comprising abdominal ultrasound images [14], are unpaired. Given this situation, we have chosen to employ CycleGAN, as it is designed for image-to-image translation tasks where paired images are unavailable. The architecture of CycleGAN includes four key components: two generator networks and two discriminator networks. The generators are responsible for translating images from one domain (e.g., CT) to another (e.g., ultrasound) and vice versa. Each generator has a corresponding discriminator that aims to distinguish between real images from the target domain and fake images created by the generator. A distinctive feature of CycleGAN is the incorporation of a cycle consistency loss. This design is based on the assumption that for instance, translating a sentence from English to French and then back to English should ideally return the original sentence, they apply a similar concept in the image-to-image translation. Mathematically, if $G : X \rightarrow Y$ represents a translator from domain $X$ to domain $Y$, and $F : Y \rightarrow X$ serves as its counterpart, then $G$ and $F$ should function as inverses to each other, with both mappings being bijections. To enforce this structure, they train the mappings $G$ and $F$ concurrently while incorporating a cycle consistency loss [17]. This loss ensures that $F(G(x)) \approx x$ and $G(F(y)) \approx y$, promoting fidelity in the translation process between the two domains. If we directly apply the CycleGAN to our task, it should follow the pipeline in Fig. 2.

### B. Proposed semantic segmentation enhanced S-CycleGAN

After training and testing an original CycleGAN model, we observed that while the overall style (color and texture) of the CT images was effectively transformed to match the ultrasound style, the anatomical details in the generated ultrasound images are hard to distinguish. This difficulty stems from the fact that in traditional image translation tasks, images from both domains are treated as samples from the joint distribution of all relevant sub-classes (such as different organs), and the translation is essentially a mapping between these distributions. Even with the use of cyclical mappings, there is no assurance that the marginal distributions of these sub-classes (or modes) are properly matched (e.g., 'liver' correctly translating to 'liver').

To maintain pixel-level semantic accuracy while converting image-level style (color and texture distribution), we incorporated two additional segmentation networks as semantic discriminators. The fake images produced by the generator are analyzed by these segmentation networks to produce a semantic mask. Subsequently, a segmentation loss is computed between this semantic mask and the label of the real source image. Moreover, unlike other similar studies [18] that continue to use an image alone as input, our network architecture employs both the image and its corresponding semantic map as inputs. This dual-input approach equips

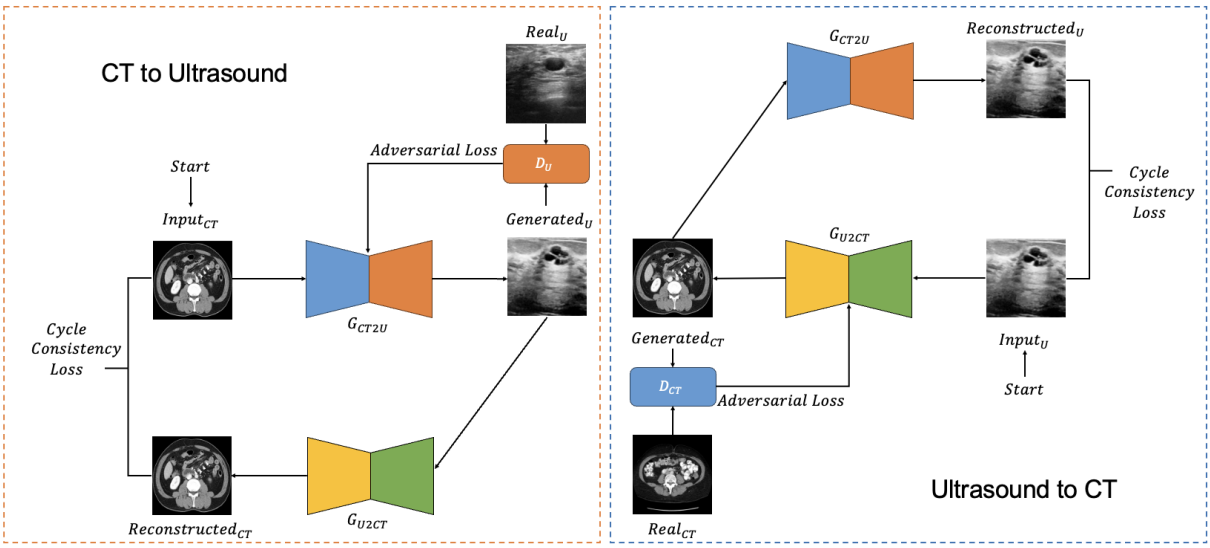
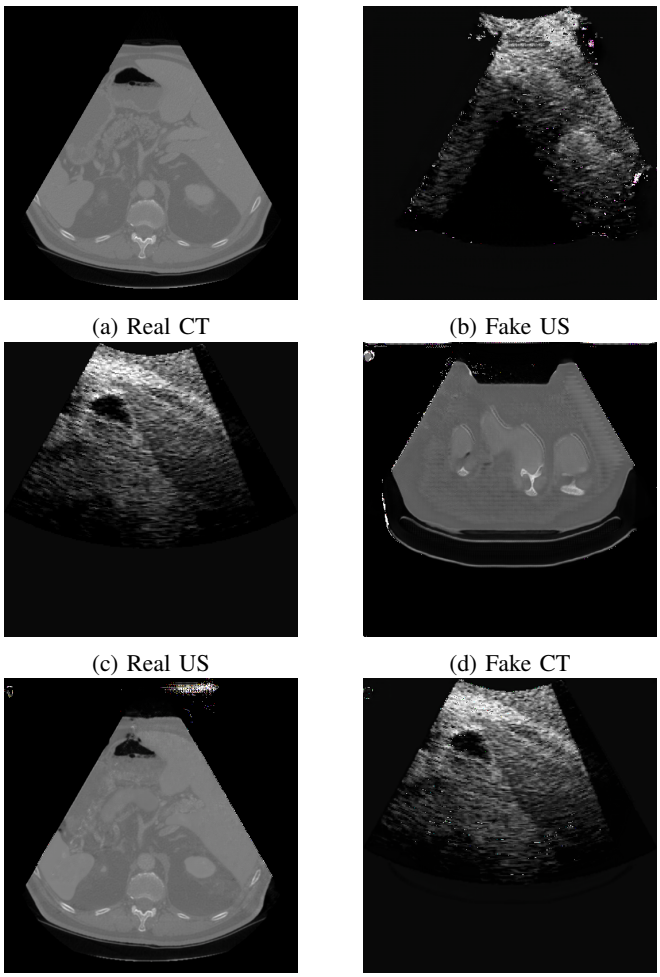



Fig. 2: CycleGAN for CT-ultrasound translation



(a) Real CT (b) Fake US

(c) Real US (d) Fake CT

(e) Reconstructed CT (f) Reconstructed US

Fig. 3: CT-ultrasound translation using pure CycleGAN

the generator with a more refined understanding of per-pixel semantic information. Hereby, we propose our S-CycleGAN (Fig. 4), which includes the following components:

*1) Generators $G_{CT\to US}$ and $G_{US\to CT}$:* These networks translate images from CT to ultrasound ($G_{CT\to US}$) and ultrasound to CT ($G_{US\to CT}$) respectively. They are trained to minimize both the adversarial and cycle consistency losses to produce realistic translations.

*2) Discriminators $D_{US}$ and $D_{CT}$:* $D_US$ discriminates between real and generated ultrasound images, whereas $D_{CT}$ differentiates real CT images from those generated by $G_{US\to CT}$. They enforce the adversarial loss component, pushing generators to create indistinguishable images from real ones.

*3) Segmentation Networks ($S_{US}$ and $S_{CT}$):* $S_{US}$ and $S_{CT}$ are responsible for generating semantic masks from ultrasound and CT images, respectively, to ensure that critical anatomical features are retained during translation.

*4) Adversarial Loss:*

$$\begin{aligned}\mathcal{L}_{adv}^{CT\to US} =& \mathbb{E}_{us\sim p_{\text{data}}(us)}[\log D_{US}(us)] \\ &+\mathbb{E}_{ct\sim p_{\text{data}}(ct)}[\log(1-D_{US}(G_{CT\to US}(ct)))]\end{aligned} \tag{1}$$

where $G_{CT\to US}$ tries to generate images that look similar to ultrasound images , while $D_{US}$ aims to distinguish between translated samples and real ultrasound image samples. $G_{CT\to US}$ aims to minimize this objective against an adversary $D_{US}$ that tries to maximize it. The counterpart is for $G_{US\to CT}$ and $D_{CT}$ is vice versa.

*5) Cycle Consistency Loss:*

$$\begin{aligned}\mathcal{L}_{cycle} =& \|G_{US\to CT}(G_{CT\to US}(x_{CT})) - x_{CT}\|_1 \\ &+\|G_{CT\to US}(G_{US\to CT}(x_{US})) - x_{US}\|_1\end{aligned} \tag{2}$$

Cycle consistency loss ensures that translating an image to the other domain and back again will yield the original image, maintaining cycle consistency across translations, which is the key to training image-to-image translation models with unpaired image sets. The L1 norm here measures the absolute differences between the original and the reconstructed image.

*6) Segmentation Loss:* The segmentation loss in the CycleGAN architecture is a combination of Cross-Entropy Loss and Dice Loss. And there are two segmentation losses computed separately on each domain: $\mathcal{L}_{seg}^{CT}$, and $\mathcal{L}_{seg}^{US}$.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{3}$$

$$\mathcal{L}_{Dice} = 1 - \sum_{i=1}^{C} \frac{2y_i \cdot p_i}{y_i + p_i + \epsilon} \tag{4}$$

where $y_i$ and $p_i$ are the groundtruth and predicted probability of being of class $i$. And $C$ is the number of classes. $\epsilon$ is an arbitrarily small smooth parameter.

The propagation flow is provided in Algorithm 1.

**Algorithm 1** S-CycleGAN Training

**Input:** Source domain $CT$, target domain $US$, hyperparameters $\lambda_{cycle}$, and $\lambda_{seg}$
Initialize networks $G_{CT\to US}$, $G_{US\to CT}$ for generation, $D_{CT}$, $D_{US}$ for discrimination, $S_{CT}$, $S_{US}$ for semantic segmentation.
**for** each epoch **do**
  **for** each batch **do**
    *Forward Pass:*
    $real\_CT, real\_CT\_Mask \leftarrow$ sample from $CT$
    $real\_US, real\_US\_Mask \leftarrow$ sample from $US$
    $fake\_US \leftarrow G_{CT\to US}(real\_CT)$
    $rec\_CT \leftarrow G_{US\to CT}(fake\_US)$
    $fake\_CT \leftarrow G_{US\to CT}(real\_US)$
    $rec\_US \leftarrow G_{CT\to US}(fake\_CT)$
    *Backward Pass and Optimization:*
    Freeze $S_{CT}$, $S_{US}$, $D_{CT}$, $D_{US}$
    Update $G_{CT\to US}$, $G_{US\to CT}$ using $\mathcal{L}_{adv}^{CT\to US} + \mathcal{L}_{adv}^{US\to CT} + \lambda_{cycle}\mathcal{L}_{cycle} + \lambda_{seg}(\mathcal{L}_{seg}^{CT} + \mathcal{L}_{seg}^{US})$
    Unfreeze $D_{US}$, $D_{CT}$
    Update $D_{US}$, $D_{CT}$ using $\mathcal{L}_{adv}^{CT\to US}$, $\mathcal{L}_{adv}^{US\to CT}$, respectively
    Freeze $G_{CT\to US}$, $G_{US\to CT}$, $D_{US}$, $D_{CT}$
    Unfreeze $S_{CT}$, $S_{US}$
    Update semantic segmentors $S_{CT}$, $S_{US}$ using $\mathcal{L}_{seg}^{CT}$, $\mathcal{L}_{seg}^{US}$, respectively
  **end for**
**end for**

## IV. EXPERIMENTS AND RESULTS

### A. Dataset

In this study, the CT data is sourced from the AbdomenCT-1K dataset [16], while the ultrasound data is obtained from the Kaggle US simulation & segmentation dataset [14]. Both datasets contain scans from the abdominal region. The CT dataset is annotated with four anatomical structures: liver, kidney, spleen, and pancreas. Conversely, the ultrasound dataset includes annotations for eight anatomical structures: liver, kidney, spleen, pancreas, vessels, adrenals, gallbladder, and bones. Therefore, for this research, we focus on the overlapping structures between the two datasets as the anatomical structures of interest. The specific organs and their corresponding mask colors are detailed in Table I.

TABLE I: Organ name and mask color

| Organ | Liver | Kidney | Spleen | Pancreas |
|---|---|---|---|---|
| Color | Violet | Yellow | Pink | Blue |

The Abdomen-1K dataset provides more than 1000 CT scans, and the data is provided in 3D format. Firstly, we randomly select 200 CT scans, and for each CT scan, we randomly sampled 10 transverse plane slices. For a more uniform image shape, we applied a fan shape mask to the CT images to mimic the outline of convex ultrasound images.

### B. Network Implementation and Training

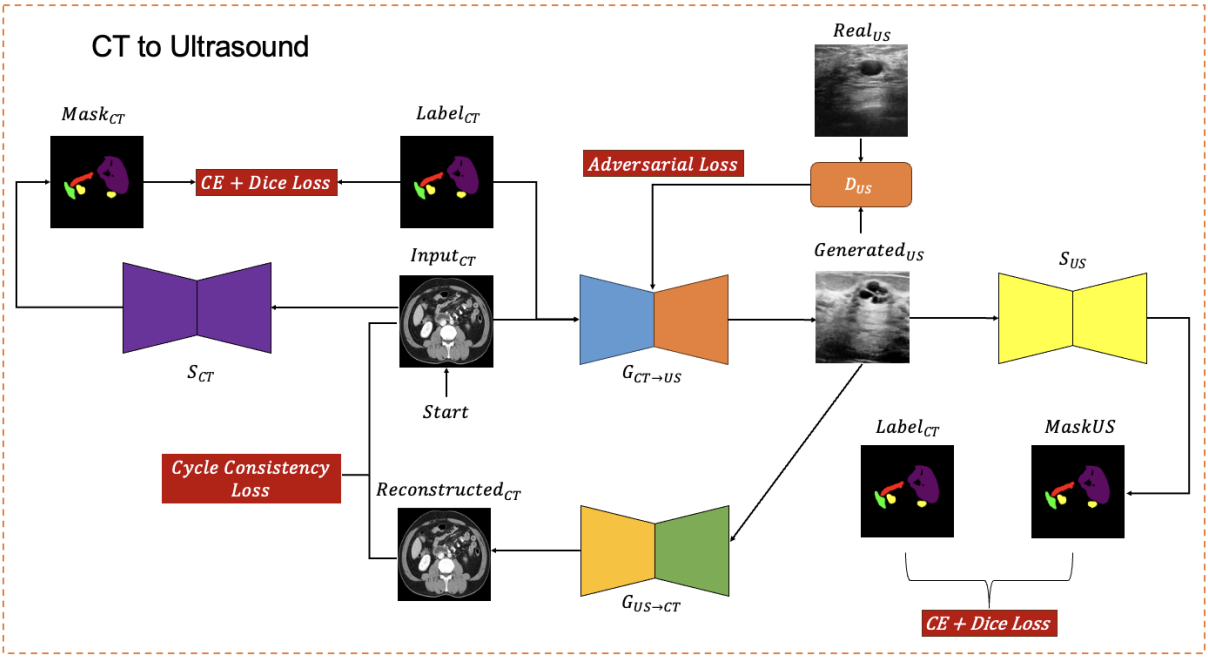
In our design, the discriminators ($D_{US}$, $D_{CT}$) follow the original design in [8]. The generators and segmentation networks are using U-Nets [19]. The generators take the concatenated image and semantic mask as input, and output a transferred image. The segmentation networks take CT or ultrasound image as input, and output predicted semantic mask. The number of convolutional filters in each U-Net block is 64, 128, and 256. The bottleneck layer has 512 convolutional filters. Initial learning rate is set as 0.0002, and will be decayed after 100 epoch with Adam. Batch size is set to 1, since we only have one GPU. The network is trained with 300 epochs in total. The coefficients $\lambda_{cycle}$, and $\lambda_{seg}$ were experimentally fixed to 10 and 0.5. One thing worth noting is that for the input of the generators and we use RGB images of CT and ultrasound (3 channels). That is for the convenience of generalizing our model to the universal tasks. The segmentation networks are trained before being incorporated into the S-CycleGAN, and using such pre-trained segmentation models help the overall network converge faster. The code implementation is using Pytorch 1.10.1, and we use one RTX 3090 Ti GPU with NVIDIA driver version 550.54.15 and CUDA version 12.4.

### C. Qualitative Results

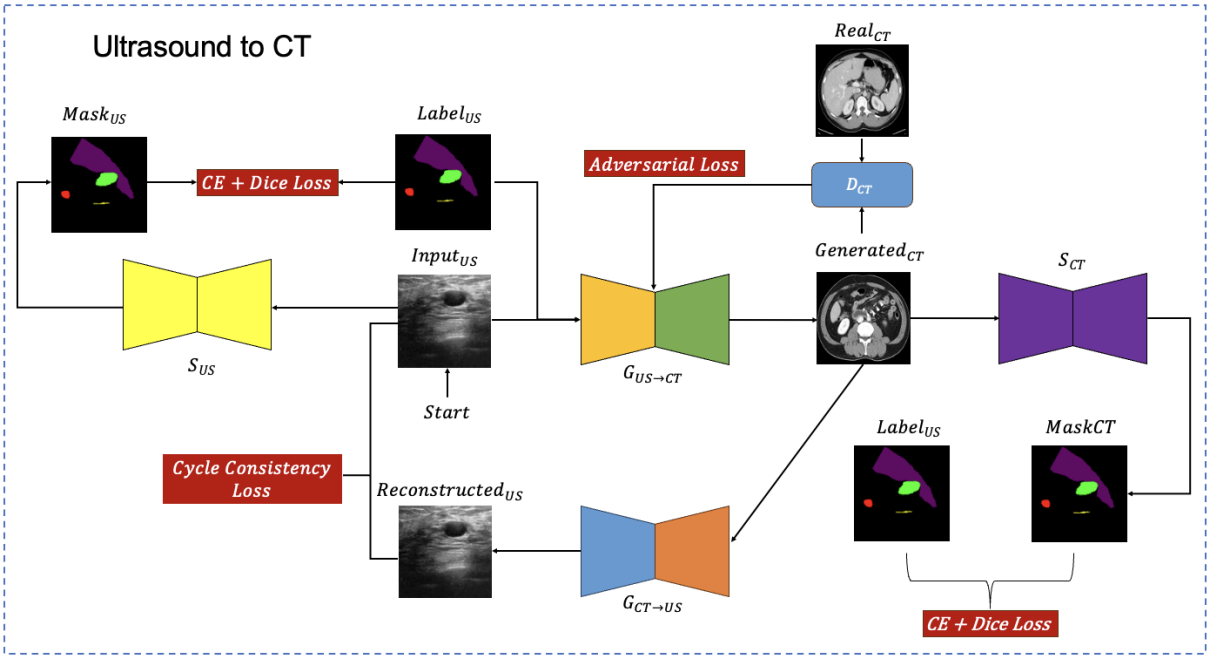
Fig. 5 and 6 present examples of the translation results from CT to ultrasound. These visual comparisons demonstrate that the S-CycleGAN can not only mimic the ultrasound style but also preserve critical anatomical features compared with Fig. 3. The synthetic images closely resemble real ultrasound scans in terms of texture and shape, suggesting a high level of detail preservation.

## V. CONCLUSION

This study introduced the S-CycleGAN, adaptation of the CycleGAN framework, enhanced with semantic discriminators for generating synthetic ultrasound images from CT data. The primary innovation of this approach lies in its ability to preserve anatomical details during the image translation process. Our model has demonstrated promising results in generating high-quality ultrasound images that closely replicate the characteristics of authentic scans. These outcomes



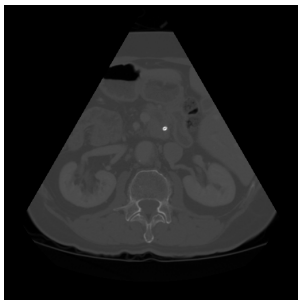

(a) CT-to-Ultrasound




(b) Ultrasound-to-CT

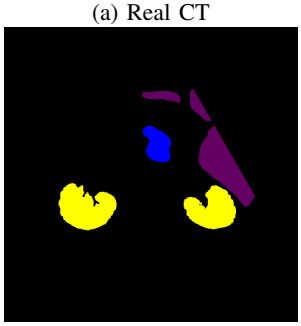
Fig. 4: Proposed Pipeline

are significant in the context of medical image translation. However, the current study is not without its limitations. Suitable metrics that comprehensively evaluate the effectiveness of ultrasound image synthesis in a numerical manner are still absent. Future work will include developing these metrics and incorporating feedback from medical experts through structured evaluation protocols. We mentioned that the synthetic dataset is expected to enhance our deep learning models. However, in current stage, the improvement by training deep learning models leveraging synthetic data is minimally significant. While our S-CycleGAN successfully replicates the visual characteristics of ultrasound images, there may still be subtle yet critical differences in textural and anatomical details compared to real ultrasound images. These discrepancies can affect the model's learning process, particularly in tasks requiring high precision. We will continue to explore additional adjustments, including the formulation of a deformation field to accurately simulate the transition from

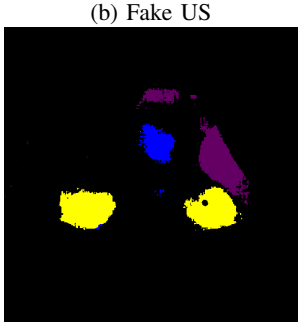


(a) Real CT

(b) Fake US

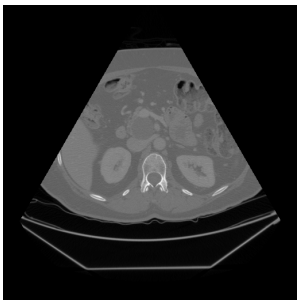


(c) CT Label

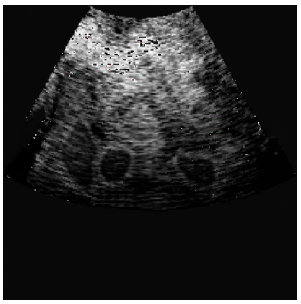


(d) Predicted US Mask

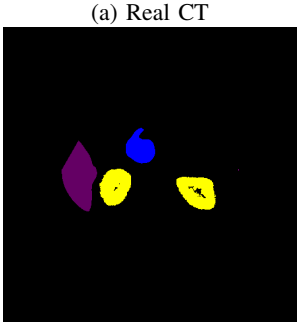
Fig. 5: CT-to-ultrasound translation example.1



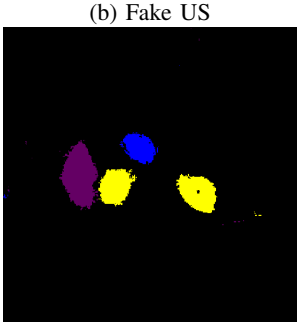
(a) Real CT



(b) Fake US



(c) CT Label



(d) Predicted US Mask

Fig. 6: CT-to-ultrasound translation example.2

CT to ultrasound imaging and a more proper training process to better leverage the synthetic data. These efforts will not only validate the clinical applicability of the synthetic images but also refine the model's performance to meet the stringent requirements of medical diagnostics.

## REFERENCES


[1] Rania Almajalid, Juan Shan, Yaodong Du, and Ming Zhang. Development of a deep-learning-based method for breast ultrasound image segmentation. In *IEEE International Conference on Machine Learning and Applications*, pages 1103–1108, 2018.

[2] S Kevin Zhou, Hayit Greenspan, and Dinggang Shen. *Deep learning for medical image analysis*. Academic Press, 2023.

[3] Johann Li, Guangming Zhu, Cong Hua, Mingtao Feng, Basheer Bennamoun, Ping Li, Xiaoyuan Lu, Juan Song, Peiyi Shen, Xu Xu, et al. A systematic collection of medical image datasets for deep learning. *ACM Computing Surveys*, 56(5):1–51, 2023.

[4] Yuhan Song, Armagan Elibol, and Nak Young Chong. Abdominal multi-organ segmentation based on feature pyramid network and spatial recurrent neural network. *IFAC-PapersOnLine*, 56(2):3001–3008, 2023.

[5] Yuhan Song, Armagan Elibol, and Nak Young Chong. Two-path augmented directional context aware ultrasound image segmentation. In *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1815–1822. IEEE, 2023.

[6] Yuhan Song, Armagan Elibol, and Nak Young Chong. Abdominal multi-organ segmentation using multi-scale and context-aware neural networks. *IFAC Journal of Systems and Control*, page 100249, 2024.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[8] 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.

[9] Bin Sun, Shuangfu Jia, Xiling Jiang, and Fucang Jia. Double u-net cyclegan for 3d mr to ct image synthesis. *International Journal of Computer Assisted Radiology and Surgery*, 18(1):149–156, 2023.

[10] Ying Chen, Hongping Lin, Wei Zhang, Wang Chen, Zonglai Zhou, Ali Asghar Heidari, Huiling Chen, and Guohui Xu. Icycle-gan: Improved cycle generative adversarial networks for liver medical image generation. *Biomedical Signal Processing and Control*, 92:106100, 2024.

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

[12] Jiamin Liang, Xin Yang, Yuhao Huang, Haoming Li, Shuangchi He, Xindi Hu, Zejian Chen, Wufeng Xue, Jun Cheng, and Dong Ni. Sketch guided and progressive growing gan for realistic and editable ultrasound image synthesis. *Medical image analysis*, 79:102461, 2022.

[13] David Stojanovski, Uxio Hermida, Pablo Lamata, Arian Beqiri, and Alberto Gomez. Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 34–43. Springer, 2023.

[14] Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192, 2020.

[15] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022.

[16] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2022.

[17] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 117–126, 2016.

[18] Anoop Cherian and Alan Sullivan. Sem-gan: Semantically-consistent image-to-image translation. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 1797–1806. IEEE, 2019.

[19] 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.