ULTRASOUND IMAGE SYNTHESIS USING GENERATIVE AI FOR LUNG CONSOLIDATION DETECTION

Yu-Cheng Chou[†1], *Gary Y. Li*[*2], *Li Chen*[2], *Mohsen Zahiri*[2], *Naveen Balaraju*[2], *Shubham Patil*[2], *Bryson Hicks*[3], *Nikolai Schnittke*[3], *David O. Kessler*[4], *Jeffrey Shupp*[5], *Maria Parker*[3], *Cristiana Baloescu*[6], *Christopher Moore*[6], *Cynthia Gregory*[3], *Kenton Gregory*[3], *Balasundar Raju*[2], *Jochen Kruecker*[2], *Alvin Chen*[2]

[1] Johns Hopkins University [2]Philips Ultrasound [3]Oregon Health & Science University
[4]Columbia University Vagelos College of Physicians and Surgeons
[5]MedStar Washington Hospital Center [6]Yale University School of Medicine

ABSTRACT

Developing reliable healthcare AI models requires training with representative and diverse data. In imbalanced datasets, model performance tends to plateau on the more prevalent classes while remaining low on less common cases. To overcome this limitation, we propose DiffUltra, the first generative AI technique capable of synthesizing realistic Lung Ultrasound (LUS) images with extensive lesion variability.

Specifically, we condition the generative AI by the introduced *Lesion-anatomy Bank*, which captures the lesion's structural and positional properties from real patient data to guide the image synthesis. We demonstrate that DiffUltra improves consolidation detection by 5.6% in AP compared to the models trained solely on real patient data. More importantly, DiffUltra increases data diversity and prevalence of rare cases, leading to a 25% AP improvement in detecting rare instances such as large lung consolidations, which make up only 10% of the dataset.

***Index Terms*—** Synthetic data training, conditional diffusion model, lung consolidation, Video object detection

1. INTRODUCTION

Recent advancements in generative models have significantly improved the data synthesis results to assist AI training [1, 2, 3, 4]. Typically, the generative models adhere to the *mask-and-paste* generation paradigm, synthesizing only the lesion area guided by pixel-level annotation of target lesion (e.g., segmentation mask) and then pasting the synthesized lesion onto a healthy background [5]. However, since the segmentation masks used as guides do not capture the internal structure and texture of the target lesion, the synthetic lesions often exhibit uniform structures, making them easily distinguishable from the surrounding tissue by the differences in boundary intensity. [6, 7].

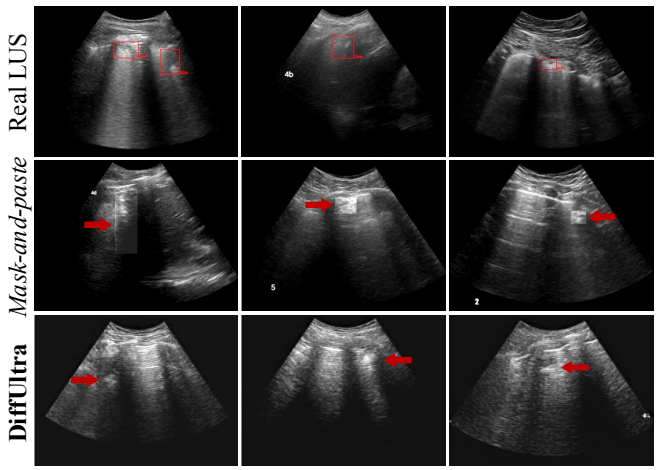

Fig. 1. Without a pixel-level lesion segmentation mask, the current *mask-and-paste* generation paradigm [6, 7] produces noticeable boundary artifacts, creating a clear distinction between the synthetic lesion and its background (middle row).

Moreover, the *mask-and-paste* generation paradigm becomes impractical when segmentation mask of a lesion is not available, rendering obvious boundary artifacts between the synthetic lesion and its background (Figure 1). Therefore, in this paper, we aim to move beyond the *mask-and-paste* generation paradigm and investigate a new approach that can synthesize structurally and positionally realistic lesions.

We hypothesize that the uniform structure of synthetic lesions stems from insufficient guidance, such as the conditions [6, 8] used in conditional diffusion models [9, 10], and poor modeling of the lesion's internal structure [1, 5, 11]. For LUS images in particular, we further hypothesize that lesion location plays a critical role in synthesizing realistic images. Specifically, as shown in Figure 3-(a), lesions of certain sizes

† Work completed during internship at Philips Ultrasound.
*Corresponding author: ye.li@philips.com

and textures can only appear on certain background patterns within the image. Placing lesions at random locations can lead to unnatural blending with the background, making the generated image appear overly artificial and potentially less effective for enhancing the downstream task.

To address this, we propose DiffUltra, a framework for synthesizing whole LUS images with lesions using structural and positional guidance. DiffUltra has two key advantages over existing methods [6, 7, 8]: (1) it models lesion-to-anatomy positions and internal structures for realistic lesion placement and texture, and (2) it generates full LUS images with only bounding box annotations, reducing annotation effort. Our experiments on a large dataset demonstrate that (1) DiffUltra produces LUS images with realistic lesion structure and position, (2) synthetic data from DiffUltra improves lung consolidation detection over models trained on real data alone(§3.2), and (3) it outperforms binary conditions (§3.3) by incorporating detailed structural representations. To our knowledge, this is the first approach to synthesize LUS images with this level of realism. Our main contributions are:

1. We introduce DiffUltra, a method for synthesizing whole LUS images with realistic lesion structure and position, without requiring pixel-level lesion segmentation.
2. We show that DiffUltra improves model reliability for lung consolidation detection compared to the model trained with real data (Table 1), especially for the rare cases (Table 2).
3. We demonstrate that conditioning the generation on structural representations enhances model performance compared to binary conditions (Table 3).
4. We show that simply duplicating rare cases does not add new information and fails to improve downstream task performance (Table 4).

2. DiffUltra

DiffUltra aims to generate realistic lesions that blend seamlessly into healthy LUS images. To place lesions in anatomically appropriate locations, we use a conditional probability mass function (PMF) based on real patient data, capturing the lesion's relative position to surrounding structures. For realistic texture, we condition the generative model with a detailed structural representation of the lesion, enabling the synthesis of whole LUS images where lesions integrate naturally with the background anatomy (Figure 3-(a)).

2.1. Lesion-anatomy Bank

2.1.1. Determining appropriate lesion position for synthesis

To ensure synthesized lesions are placed appropriately relative to their surrounding anatomical structures (e.g., the pleural line), we model the lesion's relative position to its surrounding anatomical structures in LUS images using a conditional PMF - $P(\Delta X, \Delta Y \mid X, Y)$, built from real patient data. Here, X and Y represent the coordinates of a key anatomical structure's center, while ΔX and ΔY denote the relative distance between the key anatomical structure and the lesion. This conditional PMF allows us to determine the position of the synthesized lesion by sampling from $P(\Delta X, \Delta Y \mid X = x, Y = y)$, where x and y are derived from a healthy image during synthesis.Figure 3-(a) visualizes one of these conditional PMF.

To construct the conditional PMF, we first compute the joint PMF $P(\Delta X = \Delta x, \Delta Y = \Delta y, X = x, Y = y)$ using real patient data annotated at the bounding box level. For each lesion, the distance to its nearest key anatomical structure is calculated as:

$$\begin{aligned}(\Delta x_i, \Delta y_i) &= (x' - x_i, y' - y_i),\\ (\Delta x, \Delta y) &= \min\left(\sqrt{\Delta x_i^2 + \Delta y_i^2}\right),\end{aligned} \tag{1}$$

where (x', y') and (x_i, y_i) are the bounding box centers of the lesion and its surrounding anatomical structure i, respectively. This approach allows precise modeling of relative positions, even when multiple key anatomical structures are present in the image. For each scanning zone and orientation, the joint PMF is built by counting occurrences in a 4D grid (e.g., a 10x10x10x10 grid for a given coordinate system). Next, we obtain $P(X, Y)$ by marginalizing out ΔX and ΔY from the joint PMF. The final conditional PMF $P(\Delta X, \Delta Y \mid X, Y)$ is then obtained by dividing the joint PMF by $P(X, Y)$.

2.1.2. Selecting appropriate lesion for synthesis

After determining the position of the lesion to be synthesized in the healthy image (by sampling $P(\Delta X, \Delta Y \mid X = x, Y = y)$), the next step is to select a lesion with the appropriate size and texture that fits the sampled relative position $(\Delta x, \Delta y)$. To achieve this, we propose a lesion-anatomy bank that stores lesion foregrounds (regions inside the lesion's bounding box) extracted from real patient data, indexed by $(\Delta x, \Delta y, x, y)$. During inference, a lesion foreground is randomly selected from the bank for the target position. Texture and size information are extracted from the selected foreground using Otsu' segmentation [12] (as shown in Figure 3-(b)) and used as conditions for the generative model as shown in Figure 2 (§2.3). This process is represented mathematically as:

$$P(L \mid \Delta X = \Delta x, \Delta Y = \Delta y, X = x, Y = y), \tag{2}$$

where L is the lesion foreground index. Since lesion foregrounds are unique, this conditional PMF is uniform, allowing for random sampling to retrieve a variety of lesion foregrounds for image synthesis.

2.2. Lesion Structural Representation

A straightforward method for adding synthetic lesions to a healthy image is by pasting a sampled lesion foreground onto

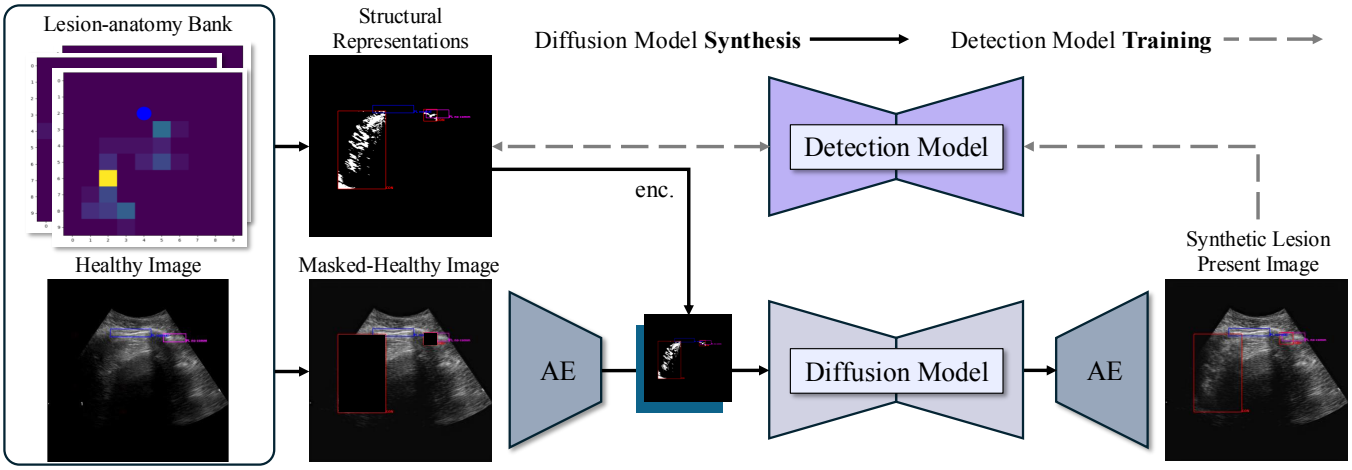

Fig. 2. The pipeline of the proposed DiffUltra.

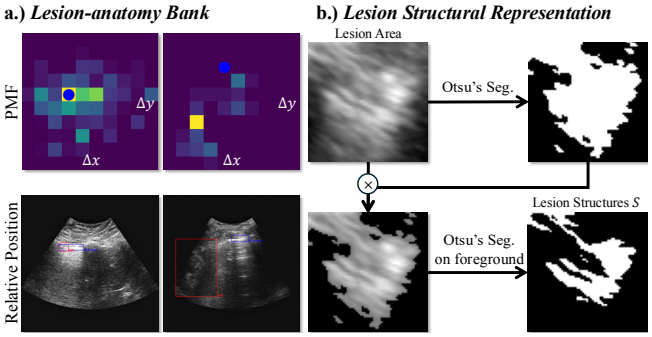

Fig. 3. a.) The visualization of $P(\Delta X, \Delta Y \mid X = x, Y = y)$, where the blue dot (x, y) denotes position of the anatomical structure, and b.) the pipeline for creating the lesion "skeleton" which is used as condition to guide the generative model.

the selected location. However, this approach lacks texture variation, as it simply copies the original lesion. To introduce more variability, we condition the generative model using a detailed structural representation of the lesion, which serves as the lesion's "skeleton", with the texture ("meat") removed. This enables the generative model to introduce texture variation during synthesis. To extract the lesion skeleton S, as illustrated in Figure 3-(b), we applied Otsu's segmentation [12] to the lesion foreground capture fine-grained structural details.

2.3. Conditional Diffusion Model

Unlike Medfusion [13], which uses covariables like age and sex, we condition the model on structural representations and latent features. Following [13], we use a stable diffusion model [9] conditioned by structural representations and latent features from a pre-trained autoencoder. With the autoencoder $Dec(Enc(\cdot))$ and diffusion model $D(\cdot)$, we generate lesions in healthy LUS images $\hat{I}$ as:

$$\hat{I} = Dec(D(f, S)), \tag{3}$$

where S is the lesion skeleton in 2.2 and f is the latent feature, obtained by:

$$f = Enc(\hat{I}_{\text{masked}}),\ \hat{I}_{\text{masked}} = \begin{cases} \hat{I}(x,y), & \text{if } (x,y) \notin F, \\ 0, & \text{if } (x,y) \in F, \end{cases} \tag{4}$$

with $\hat{I}_{\text{masked}}$ representing the masked healthy image by the entire bounding box area of the lesion foreground F, sampled from $P(L \mid \Delta X = \Delta x, \Delta Y = \Delta y, X = x, Y = y)$. The lesion's center is determined by:

$$(x', y') = (x + \Delta x, y + \Delta y). \tag{5}$$

3. EXPERIMENTS

3.1. Experimental Setting

Implementation Details. To reduce the input dimensions for the stable diffusion model, we trained an autoencoder (AE) to downsample LUS frames from $512 \times 512 \times 1$ to $64 \times 64 \times 8$ latent features. The AE was trained on all frames, and the best checkpoint was selected based on the lowest validation Mean Squred Error (MSE) loss. The diffusion model was then trained in this latent space. The diffusion models were trained for 50 epochs on 4 A100 GPUs with a batch size of 8, and the best checkpoint was chosen based on the lowest MSE loss in the foreground. During synthesis, we randomly sampled healthy images that have pleural line boxes to generate lesion-present images. Following [6], we replaced the sampled healthy images with synthetic lesion-present images to maintain a consistent number of images in the train set for detection model training. The diffusion model was set to 150

Table 1. Compared to the baseline model trained solely on real data, DiffUltra can generally improve consolidation detection performance.

	Lesion-level (AP@0.5)	Video-level (AUROC)
Yolo-v5 w/o Image Synthesis	12.7%	90.0%
Yolo-v5 + DiffTumor [6]	14.7%	91.0%
Yolo-v5 + DiffUltra	(+5.6%) **18.3%**	(+1.4%) **91.4%**

Table 2. Detection performance shows that DiffUltra can increase data diversity and thereby greatly improve the detection rate (*video-level Sensitivity*) for rare cases.

	Level 1 (**1.5%**)	Level 4 (**10.7%**)	Level 3 (15.0%)	Level 2 (23.7%)
Baseline	44.4%	42.2%	71.1%	**84.5%**
DiffUltra	(+22.3%) **66.7%**	(+25%) **67.2%**	(+2.2%) **73.3%**	82.4%

steps during inference. We used YOLO-v5 [14] and trained it for 300 epochs with default settings.

Dataset. A dataset of 7,017 LUS videos from 424 patients across 11 U.S. sites, suspected of having lung consolidation, was used. It was divided into training, validation, and testing sets, consisting of 4,930, 1,051, and 599 videos, respectively. Lung consolidation was annotated by Ultrasound-trained physicians, resulting in 45,210 bounding boxes. Additionally, 26,556 images in the training set have annotated pleural lines. Videos were sampled at every 5th frame to improve training efficiency.

3.2. Experimental Results

Synthetic Data Improves Downstream Tasks To evaluate the performance gain provided by DiffUltra, we compare a detection model trained on both real and synthetic data with one trained only on real data. Results in Table 1 show that the model trained on both real and synthetic data outperforms the baseline in lesion-level AP and video-level AUROC (+5.6% and +1.4%). Furthermore, DiffUltra outperforms DiffTumor [6] that generates lesions using the *mask-and-paste* paradigm, highlighting the effectiveness of our method in synthesizing complete LUS images.

Synthetic Data Alleviates Class Imbalance To evaluate DiffUltra's impact on improving performance in the low prevalence cases, we conducted a sub-analysis on detecting consolidations across four severity levels. To ensure a fair comparison, we matched the specificity of the baseline model (89.1%) with that of DiffUltra (88.8%). As shown in Table 2, DiffUltra significantly enhances detection rates for level-1 and level-4 consolidations (+22.3% and +25%), highlighting its effectiveness in handling rare cases, which represent only 1.5% and 10.7% of the testing set, respectively.

Table 3. Ablation results in excluding structural representation (S) and positional guidance (PMF), respectively.

S	PMF	Lesion-level (AP@0.5)	Video-level (AUROC)
		13.5%	89.8%
	✓	11.7%	**91.9%**
✓		14.8%	91.0%
✓	✓	**18.3%**	91.4%

Table 4. Hyper-parameter searching experiments of changing the positive-to-negative ratio ($P : N$) and of simply repeating rare cases ($Repeat$).

	$P : N$	Lesion-level (AP@0.5)	Video-level (AUROC)
Baseline	1: 7.6	12.7%	90.0%
$Repeat$	1: 2.3	12.7%	92.3%
DiffUltra	1: 3.3	10.6%	91.2%
DiffUltra	1: 2.5	13.0%	**92.4%**
DiffUltra	1: 1.9	**18.3%**	91.4%
DiffUltra	1: 1.5	14.6%	91.5%

3.3. Ablation Study

Excluding structural representation. Conditioning the generative model on a binary mask significantly reduces performance compared to using structural representations (12.8% vs. 18.3%, Table 3). This results are in line with our hypothesis about the need for more detailed structural representation as diffusion condition.

Excluding positional guidance. We show that randomly placing lesions creates unrealistic relations with surrounding anatomy, reducing detection performance (14.8% vs. 18.3%, Table 3). This results are in line with our hypothesis about need for realistic lesion location during synthesizing.

Repeating rare cases. We also tested whether simply repeating rare cases would improve performance by balancing the data. However, as shown in Table 4, repeating rare cases did not enhance performance, indicating that balancing alone, without new information, is insufficient for improvement.

Searching Optimal Amount of Synthetic Data We optimized the amount of synthetic data generated through experimentation. By replacing only healthy LUS images with lesion-present ones, the overall training data volume stayed constant while increasing the positive-to-negative (P) ratio. Experimental results are shown in Table 4.

4. CONCLUSION

We present DiffUltra, a method for synthesizing Lung Ultrasound images with flexible, clinically accurate lesions. Using a Lesion-Anatomy Bank, DiffUltra captures structural and positional relationships, generating realistic anatomy-lesion combinations to boost data diversity and prevalence. This approach enhances AI detection models, particularly in rare positive cases, and improves AI-based lung consolidation detection and ultrasound diagnostic reliability.

5. REFERENCES

[1] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou, "Early detection and localization of pancreatic cancer by label-free tumor synthesis," *arXiv preprint arXiv:2308.03008*, 2023.

[2] Linkai Peng, Zheyuan Zhang, Gorkem Durak, Frank H Miller, Alpay Medetalibeyoglu, Michael B Wallace, and Ulas Bagci, "Optimizing synthetic data for enhanced pancreatic tumor segmentation," *arXiv preprint arXiv:2407.19284*, 2024.

[3] Lingting Zhu, Noel Codella, Dongdong Chen, Zhenchao Jin, Lu Yuan, and Lequan Yu, "Generative enhancement for 3d medical images," *arXiv preprint arXiv:2403.12852*, 2024.

[4] 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*, vol. 79, pp. 102461, 2022.

[5] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou, "Label-free liver tumor segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7422–7432.

[6] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou, "Towards generalizable tumor synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11147–11158.

[7] Linshan Wu, Jiaxin Zhuang, Xuefeng Ni, and Hao Chen, "Freetumor: Advance tumor segmentation via large-scale tumor synthesis," *arXiv preprint arXiv:2406.01264*, 2024.

[8] Hantao Zhang, Jiancheng Yang, Shouhong Wan, and Pascal Fua, "Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models," *arXiv preprint arXiv:2403.14066*, 2024.

[9] 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*, 2022, pp. 10684–10695.

[10] 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*, 2023, pp. 3836–3847.

[11] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou, "From pixel to cancer: Cellular automata in computed tomography," *arXiv preprint arXiv:2403.06459*, 2024.

[12] Nobuyuki Otsu et al., "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[13] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, no. 1, pp. 12098, 2023.

[14] Glenn Jocher, "YOLOv5 by Ultralytics," May 2020.