ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888253

Deep Sylvester Posterior Inference for Adaptive Compressed Sensing in Ultrasound Imaging

Simon W. Penninga
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
s.w.penninga@tue.nl

Hans van Gorp
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
h.v.gorp@tue.nl

Ruud J.G. van Sloun
Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
r.j.g.v.sloun@tue.nl

***Abstract*—Ultrasound images are commonly formed by sequential acquisition of beam-steered scan-lines. Minimizing the number of required scan-lines can significantly enhance frame rate, field of view, energy efficiency, and data transfer speeds. Existing approaches typically use static subsampling schemes in combination with sparsity-based or, more recently, deep-learning-based recovery. In this work, we introduce an adaptive subsampling method that maximizes intrinsic information gain *in-situ*, employing a Sylvester Normalizing Flow encoder to infer an approximate Bayesian posterior under partial observation in real-time. Using the Bayesian posterior and a deep generative model for future observations, we determine the subsampling scheme that maximizes the mutual information between the subsampled observations, and the next frame of the video. We evaluate our approach using the EchoNet cardiac ultrasound video dataset and demonstrate that our active sampling method outperforms competitive baselines, including uniform and variable-density random sampling, as well as equidistantly spaced scan-lines, improving mean absolute reconstruction error by 15%. Moreover, posterior inference and the sampling scheme generation are performed in just 0.015 seconds (66Hz), making it fast enough for real-time 2D ultrasound imaging applications.**

***Index Terms*—active inference, cognitive systems, free energy, perception-action, ultrasound imaging, generative modelling**

I. Introduction

Ultrasound systems perform sequences of pulse-echo experiments, called transmit events, to form an image. Due to the physical speed of sound, the optimization of these transmit events constitutes a trade-off between frame rate, depth of view, and image quality, making acquisition time a major limiting resource. By reducing the amount of transmit events required to form an image, the effective budget one can spend on this trade-off improves greatly. In addition, subsampling can reduce data transfer and battery drain, enabling cheaper and more portable ultrasound systems.

Efficient subsampling and signal recovery can be achieved with Compressed Sensing [1], in which sparsity in some signal domain is used for effective reconstruction from compressed measurements, such as an undersampled Fourier spectrum and or observations from sparse arrays. Contemporary recovery methods go beyond signal sparsity and use deep learning to exploit the signal structure learned from the training data. In particular, deep generative models explicitly learn signal priors that can subsequently be used for inference. Such approaches have also been used in the context of ultrasound imaging [2, 3]. Deep learning also enables the optimization of the subsampling schemes themselves [4]. For instance, Deep Probabilistic Subsampling (DPS) [5] uses an end-to-end deep learning training method that finds the optimal subsampling strategy for downstream recovery tasks. Additional examples include subsampling of RF data through deep learning [6] and randomized channel subsampling for increased ultrasound speeds [7]. However, the aforementioned methods have in common that the learned subsampling masks are fixed, and their optimization does therefore not benefit from any information gained across the sequential sampling process at inference time. Conversely, adaptive sensing methods exploit previously acquired data to optimize future sampling schemes across a sequence of observations to improve performance [8].

In this paper, we propose an active subsampling method for ultrasound imaging that: (1) exploits a deep generative latent variable model and combines it with a deep Bayesian posterior encoder that performs fast inference of the parameters of its approximate latent posterior from partial observations; (2) designs adaptive subsampling schemes that maximize information gain *on the fly* across a sequence of ultrasound image frames in a video. Specifically, we optimize the evidence lower bound and train a deep neural network to estimate the parameters of the intricate latent posterior state distribution under partial observations, which we parameterize using a Sylvester normalizing flow [9]. Based on samples from this posterior, we subsequently design a new sampling scheme that optimizes an estimate of the expected information gain, by maximizing the marginal entropy for future observations. Both steps of the approach are illustrated in Fig. 1.

Most related to our approach, van de Camp *et al.* recently proposed the use of deep generative latent variable models for adaptive subsampling designs [10]. While effective, the method relied on Markov Chain Monte-Carlo methods for generating samples from the posterior, rendering inference prohibitively slow for time-sensitive applications such as ultrasound imaging. Moreover, the scene was considered static, and observations of this static scene were taken one at a time. In contrast, we here operate on sequences of ultrasound frames and design full subsampling masks for each next frame sequentially. Using the Sylvester normalizing flow-based posterior encoder (requiring only a single neural function evaluation), we reduce inference time by several orders of magnitude, enabling real-time processing rates, while retaining the ability to fit intricate posteriors.

The remainder of this paper is organized as follows; Section II describes the problem setup for ultrasound line-scanning, our approach to fast posterior inference, and the design of sampling schemes based on mutual information. In Section III the method is applied to sequences of ultrasound frames, and compared against non-adaptive baselines. Finally, in Section IV, we conclude and outline future work.

This work was supported by the European Research Council (ERC) under the ERC starting grant nr. 101077368 (US-ACT).

We thank SURF (www.surf.nl) for their support in using the Dutch National Supercomputer Snellius.

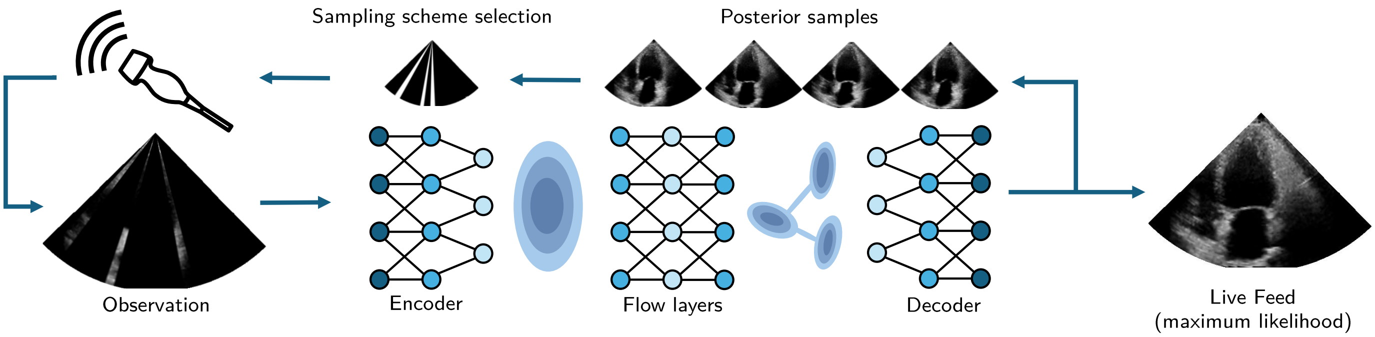

Fig. 1. Schematic overview of the active sampling loop of a single video frame. Partial observations of the full frame are used to estimate the latent posterior distribution of the next frame of the video. Samples from this posterior distribution are used to estimate mutual information between the state and the observation, which in turn determines the next subsampling mask and results in new observations.

II. Methods

A. Problem setup

Let a partial observation of a video frame at a given time-step y_t be defined as:

$$y_t = A_t x_t + n_t, \tag{1}$$

where $A_t \in \mathbb{R}^{M \times N}$ is the binary subsampling matrix (with $M << N$), $x_t \in \mathbb{R}^N$ is the fully-sampled vectorized video frame that we will refer to as the *image state*, and n_t is the added noise at the time of observation $t \in [0, 1, ..., T]$. The goal is to (1) perform efficient estimation of the Bayesian posterior for image states $p(x_t|y_t, A_t)$, and (2) design an optimal future sampling matrix A_{t+1} that maximizes expected information gain.

Unfortunately, computing the true Bayesian posterior quickly turns intractable in high dimensions. To overcome this, we use a deep latent variable model that approximates the true distribution of signals $p(x) \approx \int p_\theta(x|z)p(z)dz$ using a simplified and lower-dimensional latent distribution $p(z)$, with $z \in \mathbb{R}^{D_z}$ and $D_z << N$. Our goal then becomes to infer $p(z_t|y_t)$. When confronted with strongly subsampled image states and highly ambiguous observations, i.e. $M << N$, this posterior will nevertheless remain intricate and often multi-modal.

B. Deep Sylvester Posterior Inference

To model the complex distribution $p(z_t|y_t) \approx q_\phi(z_t|y_t)$, we use a Sylvester Normalizing Flow (Sylvester-NF). The model architecture is an extension of the Variational Auto Encoder (VAE) [11] and consists of a convolutional image encoder and decoder. The encoder q_ϕ outputs the latent Gaussian distribution parameters $\mu(y_t) \in \mathbb{R}^{D_z}$, $\sigma(y_t) \in \mathbb{R}^{D_z}$ and Normalizing Flow [12] parameters $\lambda \in \mathbb{R}^{N_p}$ for subsequent transformations of the Gaussian distribution, with N_p the parameters of the normalizing flow layers. To train the image encoder, we minimize its variational free energy. Given an observation $\hat{y}_t$ we can formulate the variational free energy as:

$$-\mathcal{F}(\theta, \phi; \hat{y}_t) = \mathbb{E}_{q_\phi(z_t^{0:K}|\hat{y}_t)}\Bigg[\underbrace{\log p_\theta(y_t = \hat{y}_t|z_t^K)}_{\text{Likelihood } Dec_\theta} - \underbrace{\log q_\phi(z_t^0|\hat{y}_t)}_{\text{Likelihood } Enc_\phi} - \underbrace{\log p(z_t^K)}_{\text{Likelihood } z_t^K} + \underbrace{\sum_{k=1}^{K} \log \left| \det \left(J[z_t^k, \lambda_k(\hat{y}_t)] \right) \right|}_{\text{LogDet Jacobian Flow transformations}}\Bigg], \tag{2}$$

in which $z_t^0 \in \mathbb{R}^{D_z}$ is a sample drawn from $\mathcal{N}(\mu(\hat{y}_t), \sigma(\hat{y}_t))$ and $z_t^K \in \mathbb{R}^{D_z}$ the same sample warped through $k \in [1, ..., K]$ flow layers. Here, J denotes the Jacobian matrix and λ_k the transform parameters of layer k. Note that we leave the dependency on the subsampling mask that generates the observations $\hat{y}_t$ implicit throughout this paper.

The generative latent variable model is first pre-trained using a dataset of full observations $\hat{x}_t \in \chi$ to capture the signal prior $\int p_\theta(x_t|z_t^K)p(z_t^K)dz_t^K$, after which the weights θ are frozen and the inference model $q_\phi(z_t^K|y_t)$ can be trained for a dataset of partial observations.

C. Sampling Policy

Our sampling policy is to maximize the information gain of future observations, which is equivalent to minimizing the expected posterior entropy [13]. The action-conditional mutual information between future latent states z_{t+1} and observations y_{t+1} for a greedy (one-step-ahead) policy is given by:

$$I(y_{t+1}; z_{t+1}|A_{t+1}, \hat{y}_t) = H(y_{t+1}|A_{t+1}, \hat{y}_t) - H(y_{t+1}|z_{t+1}, A_{t+1}). \tag{3}$$

We leave the exploration of a longer action horizon to future work. Since the entropy of expected observations y_{t+1} given z_{t+1} does not depend on A_{t+1} (it depends only on the noise n_{t+1}), our policy reduces to the maximization of the marginal entropy as:

$$A_{t+1}^* = \arg\max_{A_{t+1}} \left[H(y_{t+1}|A_{t+1}, \hat{y}_t)\right]. \tag{4}$$

The marginal entropy scales with the log determinant of the covariance matrix $\Sigma_{y_{t+1}|A_{t+1}}$, which we estimate using the generative model $p_\theta(x_t|z_t^K)$ and a sample aggregate of the posterior $q_\phi(z_{t+1}^K|\hat{y}_t)$, assuming an identity transition $z_{t+1}^K|z_t^K$:

$$\begin{aligned}
&\Sigma_{y_{t+1}|A_{t+1},\hat{y}_t} \\
&= \mathbb{E}_{q_\phi(y_{t+1}^K|\hat{y}_t, A_{t+1})}\left[(y_{t+1} - \mu_{y_{t+1}})(y_{t+1} - \mu_{y_{t+1}})^T\right] \\
&= A_{t+1}\mathbb{E}_{q_\phi(x_{t+1}^K|\hat{y}_t)}\left[(x_{t+1} - \mu_{x_{t+1}})(x_{t+1} - \mu_{x_{t+1}})^T\right] A_{t+1}^T \\
&\approx A_{t+1}\frac{1}{N_s}\sum_{i}^{N_s}(\tilde{x}_{t+1}^{(i)} - \mu_{\tilde{x}_{t+1}})(\tilde{x}_{t+1}^{(i)} - \mu_{\tilde{x}_{t+1}})^T A_{t+1}^T,
\end{aligned} \tag{5}$$

where N_S denotes the number of drawn posterior samples $\tilde{x}_{t+1}^{(i)}$. Because the action space for A_{t+1} scales with the binomial coefficient $\binom{N}{M}$, the computation is generally intractable and we instead explore only a subset of randomly selected actions S_A. We will refer to this policy as *covariance sampling*:

$$A_{t+1}^* = \arg\max_{A_{t+1} \in S_A} \log\det(\Sigma_{y_{t+1}|A_{t+1}}). \tag{6}$$

Because this strategy is very computationally expensive, we propose an alternative sampling strategy that assumes independence across the expected observations, computing the trace of the covariance matrix instead. We will refer to this policy as *trace sampling*:

$$A^*_{t+1} = \arg\max \mathrm{Tr}[\Sigma_{y_{t+1}|A_{t+1}}]. \quad (7)$$

Since this method ignores the local correlation structure of closely-spaced ultrasound scan-lines (which originates from the limited physical resolution), we explicitly prohibit the system from choosing neighbouring lines. As an example; given a set of candidate actions [5,4,6,12,11,13] sorted by covariance traces that are of decreasing order, our approach samples lines 5 and 12 due to the neighbour-exclusion of lines 4 and 6.

III. Experiments & Results

A. Experimental Setup

We evaluate our method using the EchoNet dataset [14], which contains 10,030 4-chamber cardiac ultrasound videos. Each video includes 50-250 grayscale frames with a resolution of 112×112 pixels, captured at 50 Hz. To standardize the data, the pixel values are normalized to the range [0,1], and Gaussian noise $n \sim \mathcal{N}(0, 0.02)$ is added for improved generalization. The dataset is divided into 6,986 training videos, 500 validation videos for model selection, and 500 test videos for final evaluation. The remaining 2,044 videos are excluded due to artifacts or missing data.

We convert Cartesian images into polar coordinates, with depth $r = \sqrt{x_c^2 + y_c^2}$ and scan-line angle $\gamma = \arctan\left(\frac{y_c}{x_c}\right)$. To subsample full scan-lines, A_t becomes highly structured and selects $M_\gamma < N_\gamma$ columns in the polar domain (i.e. $N = N_r N_\gamma$, $M = N_r M_\gamma$).

Our model architecture consists of a variational encoder, orthogonal Sylvester flow layers, and decoder. The encoder comprises 10 gated convolutional layers [15] with stride 2, each using $c = 64$ channels, reducing the input to a 512-dimensional latent variable $z_0 = \mu + \sigma \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, as per the reparameterization trick. The latent space is further refined using $K = 8$ normalizing flow steps, each parameterized by 16 orthogonal vectors ($N_p = 16 \times 512$), to obtain the final latent representation z^K. The decoder uses 8 blocks of gated transpose convolutions with stride 2, each using $c = 128$ channels. These layers are followed by Batch Normalization [16] and GELU activations [17]. The final image is reconstructed through a head comprising 3 additional convolutional layers, each with $c = 128$ channels.

We train the inference model with the loss function defined in (2) and we set $\beta = 1 \times 10^{-4}$ for both the generative model and the inference model. We compensate for the polar coordinate transformation by using the density of the inverse transformation as a per-pixel weighing on the training loss. In an attempt to capture all modes for a given state of observation, the IWAE [18] algorithm is used, which tightens the ELBO.

We compare the two proposed sampling policies against three baseline methods: uniform random sampling, variable density random sampling, and equispaced sampling. In uniform random sampling, independent samples from a uniform distribution are used to select the l columns for each frame. Variable density sampling uses a similar approach, but samples from a polynomial distribution centered on the middle of the image with a decay factor of 6 are used. In the equispaced policy, the system deterministically uses evenly spaced lines and shifts the set of lines by one index for each subsequent video frame, maintaining uniform sampling density across all frames. For the trace and covariance-based sampling policies, we use $N_S = 3$ posterior samples from our generative model and generate $S = 10,000$ random sampling schemes to form the candidate set S_A every t. Increasing N_S and S beyond these values resulted in increased computational costs with minimal performance improvement.

Although all subsampling methods share the same generative model, each has a distinct inference model. The training procedure is given in Algorithm 1. The computational cost of the active methods is determined by the summation of the costs of the inference, sampling, image generation, and action selection steps.

Algorithm 1 Training algorithm of the inference model for a single ultrasound video using active sampling.

Require: video $\boldsymbol{x} = [\hat{x}_1, ..., \hat{x}_T]$, empty subsampling matrix A_t, generative model parameters θ, number of posterior samples N_S and video length T.
$t \leftarrow 0$
while $t \leq T$ **do**
  $\hat{y}_t \leftarrow A_t \hat{x}_t + n_t$ ▷ Following (1)
  $z^K \leftarrow q_\phi(\hat{y}_t)$
  $\tilde{x}_t \leftarrow p_\theta(z^K)$
  $\tilde{y}_t \leftarrow A_t \tilde{x}_t$
  Back-propagate $\mathcal{L}_{\tilde{y} \rightarrow \hat{y}} + \beta \mathcal{L}_{kld(z^K)}$
  Optimizer step ϕ
  $\mu, \sigma, \lambda \leftarrow q_\phi(\hat{y}_t)$ ▷ Active sampling starts here
  Take N_S **samples** $\boldsymbol{z^K}$ **from** μ, σ, λ
  $\boldsymbol{\tilde{x}_{t+1}} \leftarrow p_\theta(\boldsymbol{z^K})$
  $A^*_{t+1} \leftarrow \Sigma_{\tilde{y}_{t+1}|A_{t+1}, \hat{y}_t}$ ▷ Following (6) or (7)
  $t \leftarrow t + 1$
end while

B. Results

Table I presents the quantitative reconstruction results, evaluated using the L1-loss, Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) across four different subsampling ratios. The proposed trace sampling method outperforms the other methods at subsampling rates of $l = 6$, $l = 9$, and $l = 15$ scan-lines. However, for $l = 12$ scan-lines, the equispaced sampling method performs slightly better. As the number of sampled lines increases, the performance gap between active and static sampling methods narrows, suggesting that active sampling is particularly advantageous when using more aggressive subsampling strategies. For equispaced and trace sampling, the upper bound for reconstruction is already approached with just $l = 15$ lines (13.4%).

A typical example (median performance) from the test set is visualized in Fig. 2, illustrating the reconstruction results for three consecutive frames for the proposed trace sampling method with $l = 6$ scan-lines. For comparison, we also present the results using equispaced sampling and full sampling, which serves as the upper (representation) limit on performance. Since all methods share the same generative model, the differences in performance can be attributed to the sampling strategies only. As seen in the absolute-difference images, in this scenario the trace sampling favoured sampling the left side of the image for t_{12} and t_{13}, leading to better reconstruction on the left at the expense of a slightly worse reconstruction on the right with respect to equispaced sampling. Interestingly, the trace-based sampling policy outperforms the covariance sampling method.

To assess the computational efficiency of our approach, we measured the time required for a complete acquisition step, including posterior estimation, on an NVIDIA GeForce RTX 2080 Ti (13.45

TABLE I

EVALUATION OF THE SAMPLING STRATEGIES FOR DIFFERENT OBSERVATION FRACTIONS. PERFORMANCE IS UPPER BOUNDED BY THE GENERATIVE MODEL, UNDER FULL OBSERVATION L1-LOSS=0.053, SSIM=0.523, PSNR=68.33.

		l = 6 (5.4% observation)			l = 9 (8.0% observation)			l = 12 (10.7% observation)			l = 15 (13.4% observation)		
	Active	L1-Loss	SSIM	PSNR	L1-Loss	SSIM	PSNR	L1-Loss	SSIM	PSNR	L1-Loss	SSIM	PSNR
Variable density	No	0.086	0.401	65.53	0.078	0.428	66.09	0.073	0.450	66.44	0.070	0.456	66.75
Uniform random	No	0.085	0.396	65.64	0.076	0.435	66.31	0.069	0.457	66.78	0.065	0.476	67.21
Equispaced	No	0.073	0.447	66.54	0.064	0.477	67.28	**0.060**	**0.494**	**67.73**	**0.058**	0.502	67.89
Covariance (ours)	Yes	0.082	0.407	65.92	0.071	0.451	66.77	0.065	0.474	67.23	0.061	0.491	67.60
Trace (ours)	Yes	**0.070**	**0.455**	**66.69**	**0.062**	**0.495**	**67.51**	0.061	0.489	67.60	**0.058**	**0.500**	**67.93**

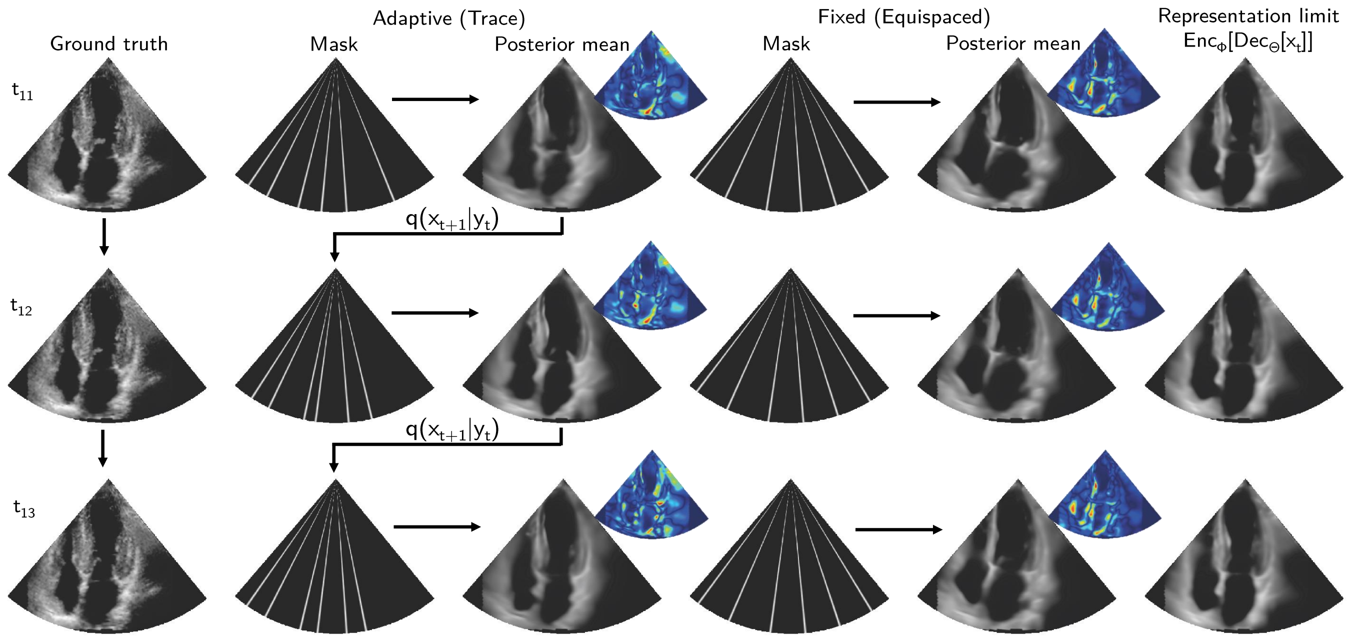

Fig. 2. Reconstruction results for three consecutive frames t_{11}, t_{12} and t_{13} of an ultrasound video that has median performance for trace-sampling (L1-Loss = 0.070). The final column shows the representation limit that is given by the deep generative model (L1-Loss = 0.055). The smaller blue cones show the absolute difference between the posterior mean and the ground truth.

TFLOPS @ FP32), using the PyTorch 2.2 [19] backend. No additional optimizations were applied, such as JIT compilation, model pruning, or quantization. For trace sampling, a single acquisition step took 0.015 seconds, while covariance sampling required 0.112 seconds. This demonstrates that the proposed approach can operate at approximately 66 Hz, or potentially faster with further optimizations, making it suitable for real-time 2D ultrasound imaging applications.

IV. DISCUSSION AND CONCLUSION

In this paper, we proposed an active subsampling method for ultrasound scan-line selection that uses an information gain maximization policy in combination with a deep generative model and a neural posterior encoder. The results demonstrate that inference can be performed successfully at subsampling rates as low as 5.4% and at frame rates of up to 66 Hz, making real-time active sampling feasible. Furthermore, we found that active sampling is especially beneficial under harsh subsampling regimes. This work opens up several avenues for future research. Firstly, because the sampling policy generates the observations on which the inference model is trained, and the inference model in turn affects the sampling policy, their optimization becomes intertwined. This may lead to collapse. In addition, the influence of the β-parameter, which determines the trade-off between accurate reconstruction and diverse samples (both affecting the accuracy of the posterior proposition), should be studied. Alternatively two separate models could be trained; one to perform accurate maximum likelihood estimation for reconstruction, and one for posterior inference, driving sampling scheme generation.

The proposed model also does not yet exploit long-term dependencies in the data, such as the cyclic nature of a beating heart. Future research to incorporate memory into the system, for example, through the use of self-attention [20] or LSTM [21], could further improve reconstruction results and/or lead to even more aggressive subsampling schemes. Additionally, the use of a more powerful deep generative model, such as the VD-VAE [22], would lead to a more accurate posterior approximation, improving the active subsampling schemes even further. Lastly, the results on 2D ultrasound show promise for application in 3D ultrasound, where the trade-off between volume rate and image resolution is far more challenging to manage.

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 4 2006.

[2] T. S. Stevens, F. C. Meral, J. Yu, I. Z. Apostolakis, J.-L. Robert, and R. J. Van Sloun, "Dehazing Ultrasound using Diffusion

Models," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2024.

[3] H. Asgariandehkordi, S. Goudarzi, M. Sharifzadeh, A. Basarab, and H. Rivaz, "Denoising Plane Wave Ultrasound Images Using Diffusion Probabilistic Models," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2024.

[4] H. Wang, E. Pérez, and F. Römer, "Deep Learning-Based Optimal Spatial Subsampling in Ultrasound Nondestructive Testing," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 9 2023, pp. 1863–1867.

[5] I. A. Huijben, B. S. Veeling, K. Janse, M. Mischi, and R. J. Van Sloun, "Learning Sub-Sampling and Signal Recovery with Applications in Ultrasound Imaging," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3955–3966, 12 2020.

[6] Y. H. Yoon, S. Khan, J. Huh, and J. C. Ye, "Efficient B-Mode Ultrasound Image Reconstruction From Sub-Sampled RF Data Using Deep Learning," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 325–336, 2 2019.

[7] J. Yu, X. Guo, S. Yan, Q. Le, V. Hingot, D. Ta, O. Couture, and K. Xu, "Randomized channel subsampling method for efficient ultrafast ultrasound imaging," *Measurement Science and Technology*, vol. 34, no. 8, p. 084005, 8 2023.

[8] H. Van Gorp, I. Huijben, B. S. Veeling, N. Pezzotti, and R. J. G. Van Sloun, "Active Deep Probabilistic Subsampling," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 8 2021, pp. 10 509–10 518.

[9] R. v. d. Berg, L. Hasenclever, J. M. Tomczak, and M. Welling, "Sylvester Normalizing Flows for Variational Inference," 3 2018.

[10] K. C. van de Camp, H. Joudeh, D. J. Antunes, and R. J. G. van Sloun, "Active Subsampling Using Deep Generative Models by Maximizing Expected Information Gain," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6 2023, pp. 1–5.

[11] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," University of Amsterdam, Amsterdam, Tech. Rep., 12 2013.

[12] D. J. Rezende and S. Mohamed, "Variational Inference with Normalizing Flows," *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538, 5 2015.

[13] D. K. Foley and E. Scharfenaker, "Bayesian Inference and the Principle of Maximum Entropy," 7 2024.

[14] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 4 2020.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language Modeling with Gated Convolutional Networks," 12 2016.

[16] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," 2 2015.

[17] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," 6 2016.

[18] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance Weighted Autoencoders," 9 2015.

[19] A. Paszke, S. Gross, F. Massa, and et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 12 2019.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Google Brain, Tech. Rep., 2017.

[21] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "xLSTM: Extended Long Short-Term Memory," 5 2024.

[22] R. Child, "Very Deep VAEs generalize Autoregressive Models and can outperform them on images," *ICLR*, pp. 1–17, 2021.