Artifact-Resilient Lung Ultrasound Feature Learning via Parameter-Efficient I-JEPA for Edge Diagnostics

Anonymous

Anonymous Institution

Abstract. Point-of-care lung ultrasound (LUS) is a portable, radiation-free modality well suited to respiratory triage, but automated interpretation is held back by scarce labels, heavy manufacturer user-interface (UI) overlays, and stochastic speckle that defeats pixel-reconstruction pretraining. We present a parameter-efficient adaptation of the Image-based Joint-Embedding Predictive Architecture (I-JEPA) for edge LUS screening. A frozen ImageNet-pretrained ViT-Small is adapted with Low-Rank Adaptation (LoRA) on its query/value projections; only the LoRA matrices and a shallow latent predictor are trained under the I-JEPA objective, which predicts target-block representations in latent space and therefore ignores speckle. On the POCOVID-Net benchmark (healthy / bacterial pneumonia / COVID-19), our model attains 0.956 ± 0.031 macro AUROC under a strict 7-fold *video-level* split while training as few as 0.037 M parameters (0.017% of the backbone). A LoRA-rank ablation shows accuracy is flat across a 16× range of adapter sizes, and representation-level probes (k-NN, silhouette, label efficiency) confirm the learned embedding is genuinely more discriminative than a frozen ImageNet baseline, most clearly when labels are scarce. The encoder runs at 29.5 ms/frame (∼34 FPS) on an Apple M4. We additionally report a negative result: under synthetic corruption the model is *not* more robust than the baseline, tempering the "artifact-resilient" claim to the preprocessing-and-latent-objective sense.

Keywords: Lung ultrasound · Self-supervised learning · I-JEPA · Parameter-efficient fine-tuning · Edge AI.

1 Introduction

Lung ultrasound has become a front-line tool for respiratory assessment because it is cheap, portable, and free of ionizing radiation [1,2]. Deploying machine learning at the point of care, however, raises three coupled obstacles. First, **labels are scarce**: curated LUS datasets are small relative to natural-image corpora. Second, **frames are contaminated by UI artifacts**—manufacturer logos, depth rulers, and text overlays—that a representation learner will exploit as shortcuts. Third, **on-device adaptation must be cheap**: a hospital should update a screening model without cloud compute, which rules out full fine-tuning of a 22 M-parameter backbone.

Self-supervised learning (SSL) addresses label scarcity, but the dominant masked-image-modelling approach reconstructs pixels [7] and thus spends capacity modelling speckle—high-frequency stochastic noise that carries no diagnostic information in ultrasound. Joint-Embedding Predictive Architectures (JEPA) [8] instead predict in latent space; I-JEPA [9] predicts the representations of target image blocks from a context block, producing semantic features without hand-crafted augmentations. This latent objective is a natural fit for ultrasound, where the signal of interest (pleural lines, consolidations, B-lines) is structural rather than textural.

We combine I-JEPA with parameter-efficient fine-tuning (PEFT). Starting from an ImageNet-pretrained ViT-Small [4], we freeze the backbone and inject LoRA [11] into the attention query/value projections, training only the LoRA matrices and a shallow predictor under the I-JEPA objective. This reduces the persisting on-device footprint to under 0.15 M parameters—and, as our ablation shows, to as few as 0.037 M with no loss of accuracy.

Contributions. (i) The first combination, to our knowledge, of a latent-predictive SSL objective (I-JEPA) with LoRA-based PEFT for lung ultrasound. (ii) A rigorous *video-level* evaluation on POCOVID-Net showing the adapted encoder beats a frozen ImageNet baseline on every clinical metric while training $< 1\%$ of parameters. (iii) A LoRA-rank ablation and three representation-level probes that explain *why* the features are better. (iv) On-device edge metrics on an Apple M4, plus an honestly reported negative robustness result.

2 Related Work

Automated Lung Ultrasound Analysis. Born et al. released POCOVID-Net and the first open-access POCUS dataset [1], and later expanded it with explainability and frame-level uncertainty, reporting COVID-19 sensitivity 0.90 ± 0.08 and specificity 0.96 ± 0.04 [2]. These works are fully supervised over the entire backbone. VanBerlo et al. showed self-supervised pretraining improves accuracy and inference efficiency across LUS tasks, especially with scarce labels [3]. We are complementary: instead of pretraining a backbone on LUS, we *adapt* a generic backbone with a self-supervised objective while updating $< 1\%$ of parameters.

Self-Supervised Visual Representation Learning. Contrastive and self-distillation methods such as BYOL [5] and DINO [6] learn invariances from augmented views; DINO further showed self-supervised ViTs [4] expose semantic attention without labels. Masked autoencoders [7] reconstruct masked pixels, but pixel-space objectives waste capacity on high-frequency detail—problematic for speckle-dominated ultrasound. JEPA [8] predicts in latent space; I-JEPA [9] predicts target-block representations from a context block, and V-JEPA [10] extends this to video. We build on I-JEPA precisely because its latent objective is expected to ignore speckle and focus on clinically meaningful structure.

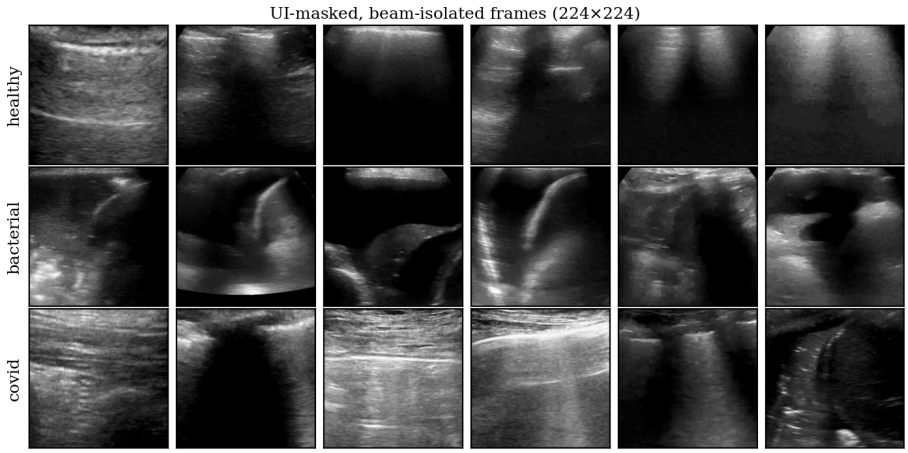

Fig. 1. Standardized POCOVID-Net frames after UI-removal masking, beam isolation, grayscale conversion and resizing to 224 × 224 (rows: healthy, bacterial pneumonia, COVID-19).

Parameter-Efficient Transfer for Medical Imaging. LoRA [11] freezes pretrained weights and injects trainable low-rank matrices into attention projections, cutting trainable parameters by orders of magnitude while matching full fine-tuning. Dutt et al. benchmarked 17 PEFT methods in medical imaging and found they match or exceed full fine-tuning, with the largest gains in low-data regimes [12]. Prior LUS work has not combined a latent-predictive SSL objective with PEFT; we close that gap.

3 Method

3.1 Data Pipeline

We use the public POCOVID-Net collection [1,2] (videos and still images, convex and linear probes), assembled by the original authors from clinical recordings collected under institutional approval and from curated public sources, and released for open research use via their public repository. As we use only this de-identified, publicly released secondary dataset, no additional ethics approval was required for this work. Frames are mapped to three classes by filename prefix (COVID-19, bacterial pneumonia, healthy; viral excluded). Each frame is converted to grayscale, a **static UI-removal mask** (configurable rectangular or radial-fan) strips manufacturer overlays and isolates the beam, the result is resized to 224 × 224 and duplicated to three channels for the ImageNet ViT. Videos are sampled every 10th frame, capped at 30 frames each, yielding **1,932 frames** (healthy 1073, bacterial 339, COVID-19 520).

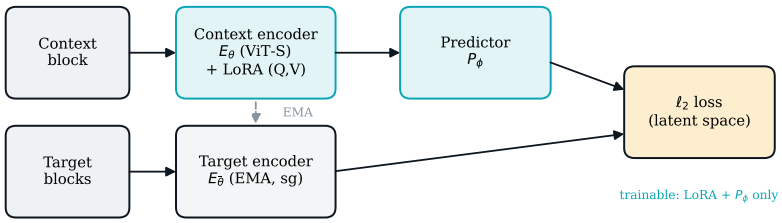

Fig. 2. Parameter-efficient I-JEPA. A context block and target blocks are encoded by a LoRA-adapted ViT-S (context) and an EMA stop-gradient copy (target); the predictor maps context tokens to target latents and the loss is computed in latent space. Only the LoRA matrices and the predictor are trained.

3.2 I-JEPA Architecture

The backbone is a ViT-S/16 (embedding dimension 384, 12 blocks, 6 heads, 196 patch tokens). I-JEPA operates on patch tokens (the class token is dropped) with three networks: a *context encoder* E_θ (the LoRA-adapted ViT) encoding a single contiguous context block (20–40% of patches); a *target encoder* $E_{\bar\theta}$, an exponential-moving-average (EMA) copy of E_θ with stop-gradient, encoding all patches, from which targets are gathered over 4 target blocks (4–10% of patches each, excluded from the context); and a *predictor* P_ϕ, a shallow 3-layer transformer (dimension 192) that takes the context tokens plus learnable mask tokens at target positions and predicts the target latents. The objective minimizes the squared ℓ_2 distance between predicted and stop-gradient target representations:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left\| \mathrm{sg}\Big(E_{\bar\theta}(x_{\text{target}}^{(i)})\Big) - P_\phi\Big(E_\theta(x_{\text{context}}),\, p^{(i)}\Big)\right\|_2^2. \tag{1}$$

3.3 Parameter-Efficient Adaptation

The ViT packs Q, K, V into one linear layer; we wrap it so frozen base outputs receive low-rank deltas added only to the **Q and V** slices (rank r, scaling α/r with $\alpha = 2r$, zero-initialized B matrices for an identity start). All backbone weights are frozen; only the LoRA matrices and the predictor are trained. At rank 8 this is 0.147 M trainable LoRA parameters (0.68% of the 21.8 M backbone); the predictor adds 1.52 M during SSL and is discarded afterwards. The persisting on-device adapter is the LoRA matrices alone.

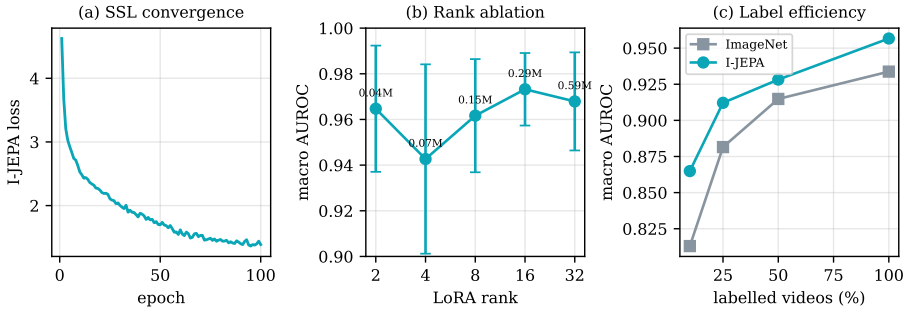

Fig. 3. (a) Self-supervised loss converges and settles without collapse. (b) Macro AUROC is flat across a 16× range of LoRA ranks (parameter counts annotated); rank 2 (0.037 M) matches rank 32. (c) I-JEPA features are more label-efficient than the frozen ImageNet baseline, with the largest margin at 10% labels.

Table 1. Main result: I-JEPA+LoRA, 7-fold video-level (mean ± std).

Metric	AUROC	Accuracy	Sensitivity	Specificity
I-JEPA + LoRA	0.956 ± 0.031	0.869	0.855	0.930

3.4 Training and Evaluation Protocol

Pretraining: 100 epochs, AdamW (lr 10^{-3}, weight decay 0.04), cosine schedule with 10-epoch warmup, batch 64, EMA momentum $0.996 \rightarrow 1.0$. After SSL we discard the predictor and target encoder, freeze the context encoder, and train a single linear layer on mean-pooled patch embeddings (class-weighted cross-entropy). **All evaluation uses a video-level split** (StratifiedGroupKFold, 7 folds): every frame of a source video stays on one side, eliminating the frame-level leakage that otherwise inflates metrics. Pretraining ran on a Modal A10G GPU (∼9 min/run); all edge timing was measured on an Apple M4 (MPS).

4 Experiments and Results

4.1 Main Result

Table 1 reports 3-class performance under 7-fold video-level evaluation. Per-class results (Table 2, representative fold) are clinically favourable, with the highest recall on COVID-19—the priority for a screening tool minimizing missed cases. The SSL loss falls ∼70% over 100 epochs and settles near 1.4 without collapsing to zero, confirming a healthy (non-degenerate) EMA/stop-gradient regime.

4.2 Does Self-Supervision Beat ImageNet?

We compare the adapted encoder against a frozen ImageNet ViT-S on identical folds (Table 3). I-JEPA wins on all four metrics. However, the AUROC gain

Table 2. Per-class clinical metrics (representative fold).

Class	Sensitivity	Specificity
COVID-19	0.986	0.964
Bacterial pneumonia	0.900	0.925
Healthy	0.835	0.956

Table 3. Frozen ImageNet baseline vs. I-JEPA+LoRA on identical folds.

Encoder	AUROC	Acc	Sens	Spec	Trainable
ImageNet baseline (frozen)	0.933 ± 0.052	0.835	0.824	0.916	0
I-JEPA + LoRA	$\mathbf{0.956 \pm 0.031}$	**0.869**	**0.855**	**0.930**	0.147 M
Δ	+0.024	+0.034	+0.031	+0.014	—

holds in only **4 of 7 folds** (Wilcoxon signed-rank $p = 0.219$) and is therefore **not statistically significant** on this small test set. The mean improvement is consistent; the representation-level evidence (Sec. 4.4) is the stronger signal.

4.3 LoRA Rank Ablation

We pretrain at LoRA ranks $\{2, 4, 8, 16, 32\}$ ($\alpha = 2r$) and probe each (Table 4). AUROC is **flat across a** $16\times$ **range of adapter sizes** (all overlap within $\pm$std). Strikingly, rank 2—37 K parameters, 0.017% of the backbone—matches rank 32. Capacity is not the bottleneck for this task, supporting an extreme-small adapter for edge deployment.

4.4 Why Are the Representations More Meaningful?

Beyond the linear probe, we examine the embedding geometry directly (frozen features, video-level folds), in Table 5. The encoder wins *without any trained classifier* (k-NN), packs classes tighter (silhouette), and helps *most when labels are scarce*—the operational signature of a genuinely better representation, and the most compelling evidence for the value of the I-JEPA objective here.

4.5 Edge-Readiness

Measured on an Apple M4 (MPS), batch 1 (Table 6). Pretraining on a Modal A10G GPU took $\sim$9 min ($\approx$5 s/epoch) versus $\sim$3.4 h on the M4 ($\sim$124 s/epoch)—a $\sim 23\times$ wall-clock speedup—while all deployment-faithful timing is reported on the M4.

4.6 Robustness (Negative Result)

We test the "artifact-resilient" hypothesis by corrupting inputs with synthetic UI blocks and speckle noise, then evaluating a probe trained on clean features (Table 7). Under this test the adapted encoder degrades *more* and ends marginally

Table 4. LoRA rank ablation, 7-fold video-level macro AUROC.

Rank	Trainable params	Macro AUROC
2	0.037 M	0.962 ± 0.019
4	0.074 M	0.946 ± 0.042
8	0.147 M	0.955 ± 0.038
16	0.295 M	0.964 ± 0.019
32	0.590 M	0.961 ± 0.025

Table 5. Representation-level probes: frozen ImageNet vs. I-JEPA+LoRA.

Probe	ImageNet	I-JEPA	Δ
k-NN AUROC (no trained head)	0.937	0.958	+0.021
Silhouette score	+0.051	+0.083	+0.032
Label efficiency @10% labels	0.813	0.865	+0.052
Label efficiency @100% labels	0.934	0.957	+0.023

below the baseline. We therefore **do not claim test-time artifact robustness**; the corruption is severe and out-of-distribution, and a milder, clinically realistic perturbation (or corruption-aware training) is needed before such a claim is warranted. The term "artifact-resilient" here refers to the static UI-removal preprocessing and the latent objective's avoidance of speckle, not to test-time corruption robustness.

5 Discussion and Limitations

The standout finding is **parameter efficiency**: a 37 K-parameter adapter matches a 590 K one, and $< 1\%$ trainable parameters beat a frozen ImageNet backbone on every clinical metric with better label efficiency and embedding geometry. This makes a strong case for on-device adaptation of LUS screening models. We are candid about three limitations. (1) The headline AUROC gain over ImageNet is *not statistically significant* ($p = 0.219$) on this small test set; larger or multi-centre data are required. (2) The *artifact-resilience claim is unsupported* by our current robustness test and is reframed accordingly. (3) Evaluation is *single-dataset* with frames from heterogeneous online sources; per-class numbers are from one fold, and external validation remains future work. Future directions include corruption-aware I-JEPA pretraining, video-level (V-JEPA-style) modelling to exploit the temporal axis of ultrasound [10], and multi-centre validation.

6 Conclusion

We presented a parameter-efficient I-JEPA for edge lung-ultrasound screening that adapts a frozen ViT-Small with LoRA, training under 0.15 M parameters

Table 6. Edge-readiness metrics (Apple M4, MPS, batch 1).

Metric	Value
LoRA trainable params	0.037 M (r2) – 0.147 M (r8)
Backbone (frozen)	21.8 M
Forward GFLOPs	∼4.3 / frame
Latency	29.5 ms / frame
Throughput	∼34 FPS (real-time)

Table 7. Synthetic-corruption robustness (clean-trained probe).

Encoder	Clean AUROC	Max corruption	Drop
ImageNet baseline	0.934	0.800	−0.135
I-JEPA + LoRA	0.955	0.787	−0.168

(as few as 0.037 M) while improving over a frozen ImageNet baseline and running in real time on an Apple M4. Representation-level probes show the learned features are more discriminative and more label-efficient. We report results under a leakage-free video-level protocol and include an honest negative robustness result, providing a realistic foundation for on-device LUS feature learning.

References

1. Born, J., Brändle, G., Cossio, M., Disdier, M., Goulet, J., Roulin, J., Wiedemann, N.: POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS). arXiv:2004.12084 (2020)
2. Born, J., Wiedemann, N., Cossio, M., et al.: Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis. Applied Sciences **11**(2), 672 (2021). https://doi.org/10.3390/app11020672
3. VanBerlo, B., et al.: Self-Supervised Pretraining Improves Performance and Inference Efficiency in Multiple Lung Ultrasound Interpretation Tasks. IEEE Access (2023). arXiv:2309.02596
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR (2021)
5. Grill, J.B., Strub, F., Altché, F., et al.: Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In: NeurIPS (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV, pp. 9650–9660 (2021)
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: CVPR, pp. 16000–16009 (2022)
8. LeCun, Y.: A Path Towards Autonomous Machine Intelligence. OpenReview, v0.9.2 (2022)
9. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In: CVPR, pp. 15619–15629 (2023)

10. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting Feature Prediction for Learning Visual Representations from Video. arXiv:2404.08471 (2024)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
12. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.: Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity. In: MIDL, PMLR **250**, pp. 406–425 (2024)