

From Points to Distributions: Probabilistic Latent Autoregression for Medical Image Representation Learning

Anonymous WACV Algorithms Track submission

Paper ID 2012

Abstract

Self-supervised learning (SSL) has fundamentally transformed medical image analysis; however, dominant pre-training paradigms rely on deterministic point estimates that ignore the inherent ambiguity of clinical data. In medical imaging, partial volume effects, subtle lesion boundaries, and acquisition artifacts create aleatoric uncertainty, which causes deterministic models to become overconfident. While the emerging paradigm of latent autoregressive models offers an efficient, decoder-free pretraining, these models still force rigid point predictions. To address this, we introduce Probabilistic Autoregression of Causal Latents (PACL), a novel SSL framework that redefines the pretraining objective from deterministic point estimation to probabilistic distributional modeling. Operating directly in the continuous embedding space without heavy pixel-level decoders or complex augmentations, PACL learns a multi-variate Gaussian distribution for the next anatomical patch conditioned on its causal context and assigns high variance to ambiguous tissue interfaces and low variance to homogeneous structures. We demonstrate that this shift to distributional learning yields highly robust, transferable representations for clinical tasks – both classification and segmentation. Extensive evaluations show that PACL outperforms previous state-of-the-art (SOTA) methods by 2% on pediatric pneumonia detection in Chest X-Rays, and by 2.1% on breast cancer detection in ultrasound. Further, PACL shows robust lesion segmentation performance on breast ultrasound and improves SOTA performance by 1%. We will release our source code and trained models.

1. Introduction

Self-supervised learning (SSL) has fundamentally transformed medical image analysis, enabling the development of highly capable visual representations without the need for exhaustive manual annotations [4, 23]. The dominant paradigms in modern visual pretraining, such as contrastive

learning [5, 8] and masked image modeling (MIM) [14, 20], focus on mapping raw pixels to a fixed-dimensional representation. While these approaches have achieved remarkable success in natural image domains, they are overwhelmingly deterministic [10]. Such methods train models to output a single, fixed feature vector for any given input, forcing a point estimate regardless of the underlying data distribution.

On the other hand, medical imaging is uniquely characterized by inherent ambiguity, where partial volume effects, subtle lesion boundaries, acquisition artifacts, and variable tissue contrast create aleatoric uncertainty [13, 40, 44]. When standard self-supervised objectives force a model to make a deterministic point estimate of an ambiguous anatomical structure, they encourage overconfidence and fail to capture the rich, probabilistic nature of clinical data [16, 17].

Recent works in the broader vision community, such as iGPT [7], JEPa [3], NEPA [52], have begun exploring a paradigm shift from traditional representation learning to predictive autoregression in the latent space. These frameworks learn to predict future patch embeddings conditioned strictly on previous ones. By eliminating pixel-level decoders, latent autoregression offers a highly streamlined and scalable pretraining alternative. Inspired by such explorations in natural vision, we investigate the applicability of latent autoregression for medical imaging. However, we identify a fundamental limitation in these existing formulations: they optimize for deterministic point estimates to a single target. Consequently, while architecturally elegant, these methods remain blind to the distributional uncertainty that is ubiquitous in clinical scans.

To address the inherent aleatoric uncertainty of clinical scans where tissue boundaries and acquisition artifacts create genuine ambiguity, we propose a conceptual leap from point estimation to distributional modeling. We introduce Probabilistic Autoregression of Causal Latents (PACL), a novel autoregressive framework that redefines the pretraining objective from point estimation to distributional model-

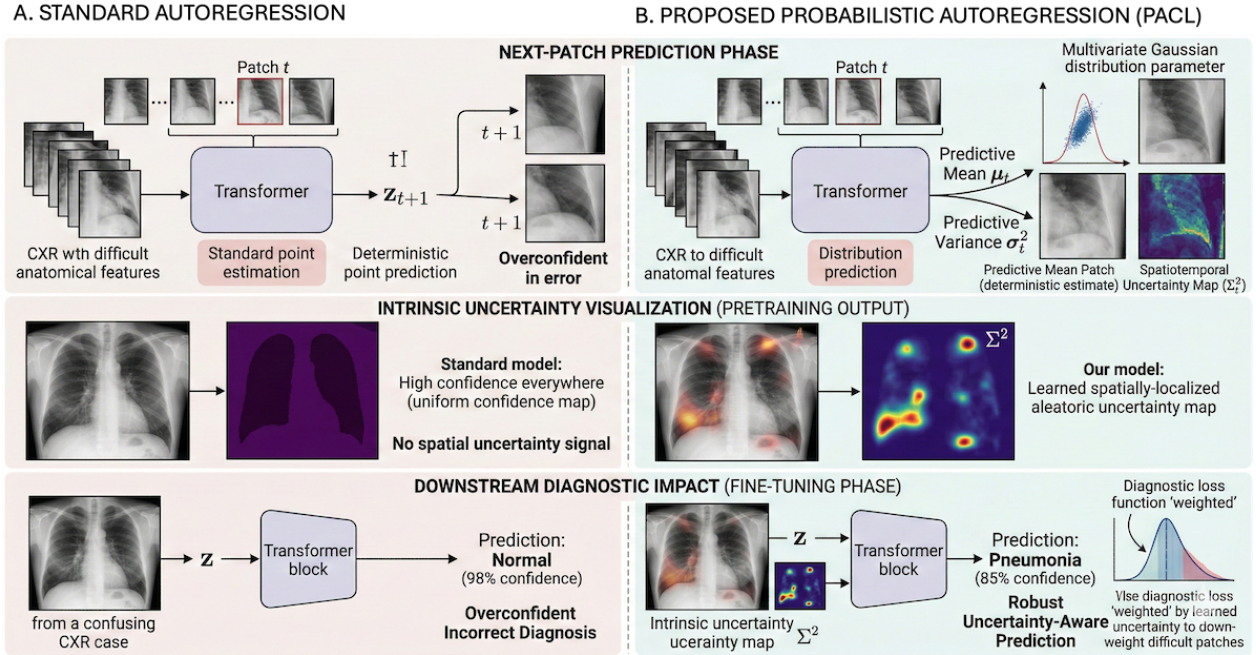


Figure 1. Motivation for PACL. (A) Standard latent autoregression relies on deterministic point estimation, which ignores spatial ambiguity and risks overconfident, incorrect predictions. (B) In contrast, PACL predicts a multivariate Gaussian for the next anatomical patch. This yields spatially-localized aleatoric uncertainty maps. During fine-tuning, the uncertainty guides the downstream loss, enabling robust clinical predictions.

ing. Unlike traditional generative pretraining methods that rely on pixel-level reconstruction or discrete tokenization, PACL treats autoregressive prediction itself as the primary learning signal, operating directly in the continuous embedding space leading to a decoder-free design. Further, PACL moves beyond the limitations of standard latent autoregression, and fundamentally alters the predictive objective. Instead of generating a deterministic point estimate for the next patch, PACL learns to output a multivariate Gaussian distribution conditioned on the causal context of the anatomy. By optimizing a negative log-likelihood (NLL) objective in the latent space, the network explicitly learns to quantify spatial uncertainty. Consequently, the model naturally assigns high variance to ambiguous tissue interfaces and acquisition artifacts, while confidently predicting low variance for homogeneous anatomical structures.

We demonstrate that predicting distributions rather than points yields robust, transferable representations for downstream medical diagnostic tasks. We validate PACL on two distinct medical image analysis tasks: Pneumonia detection in Chest X-Ray, and Breast Cancer detection in Ultrasound. We further validate the applicability of the method on dense prediction tasks by evaluating the performance on lesion segmentation from breast ultrasound data.

Contributions: The key contributions of this work are:

- We propose PACL, shifting visual autoregression from

deterministic point-estimation to probabilistic distributional modeling in the latent space, which is naturally suitable for the nuances of medical imaging.

- PACL eliminates heavy decoders and complex augmentations from its design, and preserves architectural simplicity while uniquely supporting probabilistic distributional learning for clinical tasks.
- We show the efficacy of PACL, and outperform previous SOTA on two downstream medical image classification tasks: (1) pediatric pneumonia detection in Chest X-Ray by 2%, (2) breast cancer detection in ultrasound by 2.1%.
- We further show the broader applicability of PACL on lesion segmentation in breast ultrasound to ensure that PACL captures spatially meaningful lesion structure and boundaries, not just image-level classification cues.

2. Related Work

Self-Supervised Medical Representation Learning.

Self-supervised learning (SSL) has become a central strategy for leveraging unlabeled clinical data, with contrastive and multi-instance objectives improving transfer in dermatology, radiography, and ultrasound [4, 5, 8, 23]. A second line of work formulates pretraining as reconstruction or masked prediction. Model Genesis learns 3D anatomical priors through restoration tasks [56], while MAE and OmniMAE scale masked image modeling to ViT through pixel

or patch level reconstruction [14, 20]. Medical adaptations further show that clinical recognition often depends on preserving fine texturing and morphology, for example, GLCM-MAE replaces purely pixel-wise reconstruction with a texture-aware objective tailored to medical images [31]. Task specific systems, including chest X-ray prompt learning and anomaly detection methods, have improved individual diagnostic settings [27, 46, 48, 54]. However, most SSL methods still map each image or patch to a deterministic feature, collapsing ambiguous anatomy, weak lesions, and acquisition artifacts into a single target representation rather than modeled as uncertainty.

Latent predictive and autoregressive visual learning. Autoregressive visual pretraining was first explored by modeling raw pixels or discrete visual tokens [7]. More recent joint-embedding predictive methods avoid expensive pixel decoders by predicting latent targets from visible context. I-JEPA predicts target-block representations from contextual image blocks without handcrafted augmentations [3], while NEPA formulates vision pretraining as causal next-embedding prediction using stop-gradient targets and decoder-free objective [52]. This latent predictive paradigm is attractive for medical imaging because it avoids fragile augmentations and unnecessary pixel synthesis. Nevertheless, existing formulations still optimize a point estimate for every predicted patch. PACL is closest to NEPA in its use of causal latent autoregression however, it changes the predictive family and the training objective. Whereas NEPA regresses a deterministic estimate of the next latent embedding, PACL predicts a conditional Gaussian distribution, and optimizes a heteroscedastic likelihood. This is because, in clinical scans, where partial-volume effects, low contrast, and boundary ambiguity make multiple latent explanations plausible, deterministic next-embedding prediction provides no mechanisms to represent aleatoric uncertainty.

Uncertainty and Dense clinical prediction. Uncertainty estimation has been extensively studied in medical image analysis, including Bayesian and heteroscedastic formulations for modeling epistemic or aleatoric uncertainty [13, 25, 40]. Recent studies also examine reliability metrics and probabilistic embeddings for medical models [16, 17]. However, heteroscedastic training can become unstable when high predicted variance suppresses gradients, motivating calibrated objectives such as beta-NLL [42]. In parallel, dense prediction architectures such as U-Net, nnU-Net, UNETR, Swin UNETR, and MedSAM demonstrate the importance of spatially faithful encoders for dense anatomical localization across modalities [18, 24, 30, 37, 47]. These methods, however, typically introduce uncertainty during supervised downstream prediction or model calibra-

tion. PACL instead moves uncertainty estimation into SSL pretraining itself, learning distributional latent representations that can be transferred to downstream classification and dense clinical prediction.

3. Our Method

3.1. Architecture

Causal Vision Transformer (ViT) Backbone: To enable autoregressive representation learning, we adapt a standard ViT backbone [12] with causal attention masking [7]. Given an input image partitioned into a sequence of N tokens each of D -dimension, we employ a strict causal mask on the self-attention. This guarantees that the dense hidden state extracted at step t , denoted as $\mathbf{z}_t$, is conditioned entirely on the prefix sequence $\mathbf{z}_{\leq t}$, physically preventing any information leakage from future anatomical patches. To further enhance spatial reasoning, improve training stability, prevent attention explosion, and enhance architectural robustness, we augment the backbone with Rotary Position Embeddings (RoPE) [45], LayerScale [49], QK-Norm [21], and SwiGLU [43] activations.

Probabilistic Projection Head: Moving beyond standard deterministic autoregression, which uses the hidden state $\mathbf{z}_t$ as a rigid point estimate for the next embedding $\mathbf{z}_{t+1}$, PACL models the conditional generative process as a multivariate diagonal Gaussian:

$$p(\mathbf{z}_{t+1} | \mathbf{z}_{\leq t}) = \mathcal{N}(\boldsymbol{\mu}_t, \text{diag}(\boldsymbol{\sigma}_t^2)). \quad (1)$$

To parameterize this, a single linear projection head, $h_{prob} : \mathbb{R}^D \rightarrow \mathbb{R}^{2D}$, expands $\mathbf{z}_{\leq t}$ and partitions it along the feature dimension to independently yield the predictive mean and log-variance: $(\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2) = h_{prob}(\mathbf{z}_t)$. Finally, to align with the target embedding space, $\boldsymbol{\mu}_t$ is normalized, while $\log \boldsymbol{\sigma}_t^2$ is retained to capture the spatially localized aleatoric uncertainty of the anatomical region.

3.2. Probabilistic Autoregressive Pretraining

To safely transition a ViT to stable, well-calibrated probabilistic forecasting prior to diagnostic supervision, we employ a progressive pretraining strategy. Initializing from a standard ImageNet-1k pretrained ViT, we first strictly freeze the backbone and optimize only the probabilistic projection head via Gaussian NLL on ImageNet-1k. The head forecasts next-token distributional parameters from the prefix sequence $\mathbf{z}_{\leq t}$, targeting the corresponding continuous embedding $\hat{\mathbf{z}}_{t+1}$ obtained directly from the frozen ViT encoder. To prevent representation collapse and shield the pretrained visual semantics from unstable initial gradients, a stop-gradient operation, $\text{sg}(\cdot)$, is applied to this target. Once the projection head stabilizes, both components are unfrozen for joint autoregressive optimization. By applying

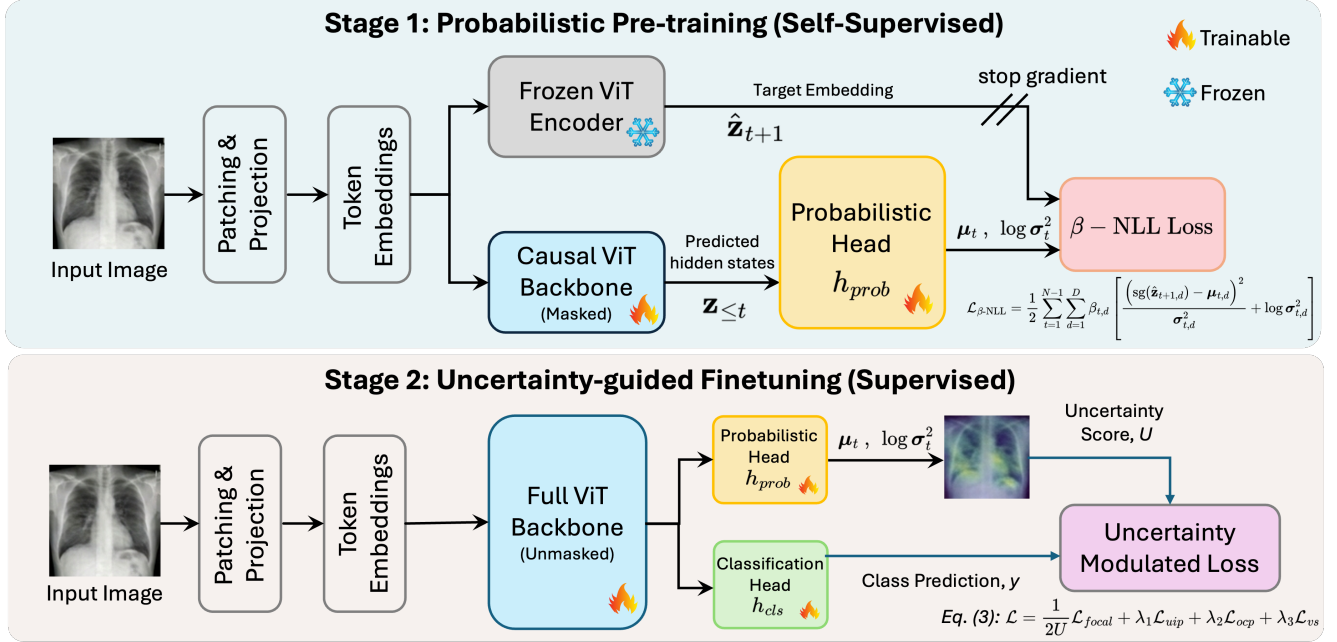


Figure 2. Overview of PACL. *Pre-training*: A masked causal ViT with a distributional head predicts latent distributions of future patches, optimized via β -NLL against frozen target embeddings (with stop-gradient). *Fine-tuning*: The unmasked ViT extracts global context. Predicted spatial variance is aggregated into an uncertainty score U to modulate ambiguity in the downstream loss. For downstream segmentation, we replace the classification head with a lightweight decoder, and use an uncertainty-modulated BCE and Dice Loss discussed in Section 3.4.

a highly constrained, differential learning rate to the backbone, the model gently adapts to distributional prediction without destroying its deterministic structural priors. We explicitly note that the ImageNet-1K pretraining is just an one-time task for stabilizing the model, and need not be repeated for any pretraining-finetuning phase for subsequent downstream tasks.

Adapting Probabilistic Pretraining on Medical Data:

We subsequently adapt the model to the unique noise profiles of target-domain medical scans. For our experiments, we used the CXR and BUSI training data, respectively for this stage. Standard Gaussian NLL aggressively down-weights high-variance tokens, potentially causing the model to ignore difficult clinical regions. To prevent this, we employ a β -weighted negative log-likelihood (Beta-NLL) inspired from [42]. By scaling the loss by $\beta_t = (\sigma_t^2)^\gamma$ (where $\gamma > 0$), we ensure that regions exhibiting severe aleatoric uncertainty still contribute meaningfully to the gradient. We formulate the objective as:

$$\mathcal{L}_{\beta\text{-NLL}} = \frac{1}{2} \sum_{t=1}^{N-1} \sum_{d=1}^D \beta_{t,d} \left[\frac{(\text{sg}(\hat{\mathbf{z}}_{t+1,d}) - \mu_{t,d})^2}{\sigma_{t,d}^2} + \log \sigma_{t,d}^2 \right],$$

$$\beta_t = (\sigma_t^2)^\gamma. (2)$$

Adapting to the medical domain risks catastrophic for-

getting of generic structural priors. To mitigate this, we use an Exponential Moving Average (EMA) update of the backbone weights during this adaptation phase. Additionally, to prevent pathological variance explosion or collapse during this domain shift, the predicted log-variance is explicitly bounded: $\log \sigma_t^2 \in [\log \sigma_{\min}^2, \log \sigma_{\max}^2]$. Through this unified training stage, the encoder successfully learns *domain-aware probabilistic representations* that encode both deep semantic structure and spatial aleatoric uncertainty.

3.3. Adapting to Downstream Classification Task

After probabilistic autoregressive pretraining, we fine-tune the model for downstream pathology detection within a Multi-Task Learning (MTL) framework. At this stage, we relax the strict causal attention of pretraining to full self-attention, allowing the network to leverage the global spatial context simultaneously. A diagnostic classifier h_{cls} , implemented as a lightweight residual bottleneck network, is attached in parallel with the probabilistic head (h_{prob}). During finetuning, h_{cls} produces class probabilities $\hat{y}$, and the probability head, h_{prob} , produces the spatial variance matrix $\Sigma^2 \in \mathbb{R}^{N \times D}$, where N is the number of image patches and D is the embedding dimension. We combine these variances in a single image-level uncertainty score which is used to formulate the proposed uncertainty-guided loss.

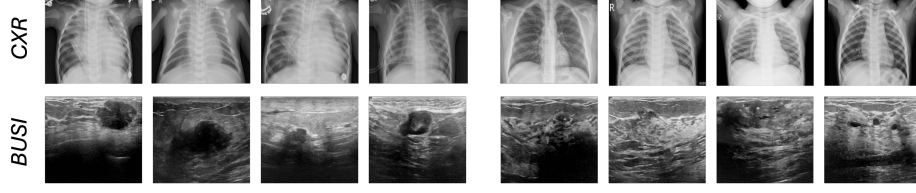


Figure 3. Sample images from the CXR [26] (left: pneumonia, right: normal) and BUSI [2] (left: malignant, right: benign) datasets.

The aggregated image-level uncertainty score is defined as:

$$U = \frac{1}{N \times D} \sum_{i=1}^N \sum_{j=1}^D \sigma_{i,j}^2. \text{ We initially freeze the backbone}$$

and h_{prob} , and warm up h_{cls} . Later, we unfreeze all modules with a joint end-to-end fine-tuning of h_{prob} and h_{cls} . Implementation details are in Sec. 5.

Uncertainty-Guided Downstream Classification Loss:

To ensure the classifier does not overfit to ambiguous or noisy clinical scans, we formulate an uncertainty-aware objective that modulates the standard classification losses based on the predicted ambiguity. The modified classification loss is inspired from [25], and is defined as: $\mathcal{L}_{cls} = \frac{1}{2U} \cdot \mathcal{L}_{base}$, where $\mathcal{L}_{base}$ denotes the base classification loss (Focal Loss in our design). The attenuation factor $\frac{1}{2U}$ explicitly controls the gradient contribution of highly uncertain scans, preventing the network from confidently learning erroneous correlations from acquisition artifacts. We clip the attenuation factor to 5, to maintain numerical stability. We penalize the unbounded growth of uncertainty by $\mathcal{L}_{uip} = \log(1 + U)$. We enforce a minimum uncertainty lower-bound (τ) with the regularization, $\mathcal{L}_{ocp} = \text{ReLU}(\tau - U)^2$, to prevent pathological overconfidence. We further stabilize the training by penalizing the high-variance uncertainty predictions by adding the loss term $\mathcal{L}_{vs} = U^2$. These regularization components are governed by weights λ_1 , λ_2 , and λ_3 . The overall uncertainty-guided objective is formalized as Eq. (3):

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{uip} + \lambda_2 \mathcal{L}_{ocp} + \lambda_3 \mathcal{L}_{vs} \quad (3)$$

3.4. Adapting to Downstream Segmentation Task

We used the probabilistic pretrained model with the ViT backbone and the probabilistic head, and added a lightweight decoder suitable for the segmentation task. The decoder converts ViT patch tokens into a binary lesion mask. The model is trained using BCE plus Dice loss with PACL uncertainty-guided weighting. Specifically, for prediction logits $\hat{y}$ and ground-truth mask y , the base segmentation loss is:

$$\mathcal{L}_{seg} = \lambda_b \mathcal{L}_{bce}(\hat{y}, y) + \lambda_d \mathcal{L}_{dice}(\hat{y}, y), \quad (4)$$

where we use $(\lambda_b, \lambda_d) = (1.0, 2.0)$. The final loss is uncertainty-weighted as:

$$\mathcal{L} = \frac{1}{2\sigma^2} \mathcal{L}_{seg} + \alpha \log(1 + \sigma^2), \quad (5)$$

where σ^2 is estimated from the PACL probabilistic head and $\alpha = 0.1$.

4. Datasets

Diagnosis of Pediatric Pneumonia in Chest X-Ray Images:

We evaluated our proposed PACL using the publicly released pediatric chest X-ray (CXR) dataset [26]. Following the official dataset split, the training set consists of 5,232 images, comprising 1,349 normal scans and 3,883 pneumonia cases. The test set includes 624 images from a distinct patient group, containing 234 normal controls along with 390 pneumonia scans.

Breast Cancer Detection in Ultrasound Images:

We further validated the applicability of our approach on breast cancer detection from ultrasound using the publicly accessible BUSI dataset [2]. The cohort comprises 210 malignant, 487 benign, and 133 normal ultrasound image samples.

5. Experiments and Results

5.1. Implementation and Evaluation

Pretraining with PACL: We use a machine with Intel Xeon Gold 5218@2.30GHz processor and 4 Nvidia Tesla V100 GPUs for our experiments. First, we use a frozen ImageNet-1k pretrained ViT backbone. The probabilistic head is trained with learning rate 10^{-3} to map semantic features into a calibrated Gaussian parameter space. Next, the full network is unfrozen and optimized using Beta-NLL loss defined in Sec. 3.2 with $\gamma = 0.5$. We use AdamW optimizer with β values (0.9, 0.999) and learning rate 5×10^{-6} for the backbone and 10^{-4} for the probabilistic head, while additionally enforcing strict log-variance bounds of $[-5, 5]$ and maintaining an EMA with decay = 0.999 of the model weights to prevent catastrophic forgetting of foundational priors.

Fine-tuning for Downstream Segmentation on BUSI:

For BUSI lesion segmentation, we add a lightweight convolutional decoder on top of the Stage-3 PACL ViT backbone. Images are resized with aspect-ratio preservation and center-cropped to 224×224 . Training uses only random zoom augmentation with scale sampled from $[0.97, 1.05]$. Validation and testing use only resize and center crop. With patch size 14×14 , the ViT produces a 16×16 token grid,

Group	Method	AUROC	Acc.	F1
Task-specific SOTA	VAE [28]	0.618	0.640	0.774
	Ganomaly [1]	0.780	0.700	0.790
	f-AnoGAN [41]	0.755	0.740	0.810
	MemAE [15]	0.778	0.565	0.826
	MNAD [34]	0.773	0.736	0.793
	SALAD [53]	0.827	0.759	0.821
	CutPaste [29]	0.736	0.640	0.723
	PANDA [36]	0.657	0.654	0.663
	M-KD [39]	0.741	0.691	0.623
	IF 2D [32]	0.810	0.764	0.822
	PaDiM [11]	0.714	0.729	0.807
	IGD [9]	0.734	0.740	0.809
	SQUID [51]	0.876	0.803	0.847
	CheXzero [48]	0.927	0.830	0.875
	Xplainer [35]	0.899	0.782	0.850
	CoOp [54]	0.946	0.846	0.886
	MaPLe [27]	0.951	0.894	0.895
	PPAD [46]	0.967	0.894	0.918
MIM Variants	MAE [20]	0.941	0.949	0.922
	OmniMAE [14]	0.953	0.950	0.968
	GLCM-MAE [31]	0.964	0.955	0.964
Autoregression Variants	NEPA [52]	0.890	0.811	0.847
	PACL (Ours)	0.987	0.955	0.965

Table 1. The AUROC, Accuracy, and F1 of SOTA baselines and PACL for identifying Pneumonia from chest X-Ray images. We report the metrics on the test set provided by original authors which consists of 390 pneumonia and 234 normal images.

Group	Method	Spec.	Sens.	AUROC
Prev. SOTA	ResNet50 [19]	0.975±0.024	0.829±0.088	-
	DenseNet121 [22]	0.968±0.018	0.824±0.027	-
	Radformer [6]	0.881±0.069	0.738±0.105	-
MIM Variant	MAE [20]	0.945±0.019	0.748±0.059	-
	OmniMAE [14]	0.947±0.034	0.614±0.124	-
	GLCM-MAE [31]	0.936±0.029	0.852±0.028	0.921±0.022
Autoregression Variant	PACL (Ours)	0.906±0.023	0.873±0.038	0.938±0.036

Table 2. The 5-fold cross validation (Mean±SD) specificity, sensitivity, and AUROC metrics of SOTA baselines and PACL in detecting breast cancer from ultrasound data.

which is reshaped into a feature map and upsampled by the decoder to predict a binary lesion mask.

The ViT backbone and probabilistic head are first frozen while the decoder is warmed up for 8 epochs with learning rate 5×10^{-4} . Then, the backbone and decoder are fine-tuned for 100 epochs, keeping the probabilistic head frozen for uncertainty estimation only. We use batch sizes 64/128 for training/validation-test, weight decay 0.01, gradient clipping at norm 1.0, learning rates 5×10^{-6} for the

backbone and 5×10^{-5} for the decoder, and select checkpoints by validation IoU (intersection-over-union). Dice and IoU are computed from masks thresholded at 0.5, and results are averaged over test images. Benign and malignant BUSI lesions are both treated as foreground, yielding binary lesion-versus-background segmentation.

Evaluation Metrics: The proposed framework is evaluated on AUROC, Accuracy, and F1 metrics for pediatric pneumonia detection on Chest X-Ray using the official train-

Method	IoU	Dice
U-Net[38]	0.572±0.047	0.719±0.035
Att-UNet[33]	0.552±0.036	0.702±0.029
U-Net++ [55]	0.574±0.048	0.721±0.039
U-Next [50]	0.591±0.010	0.731±0.013
PACL (Ours)	0.604±0.019	0.706±0.020

Table 3. The IoU and Dice scores of SOTA baselines and PACL for lesion segmentation on breast ultrasound data [2]. Mean and standard deviation over three seeds are reported for a stratified 80:20 split.

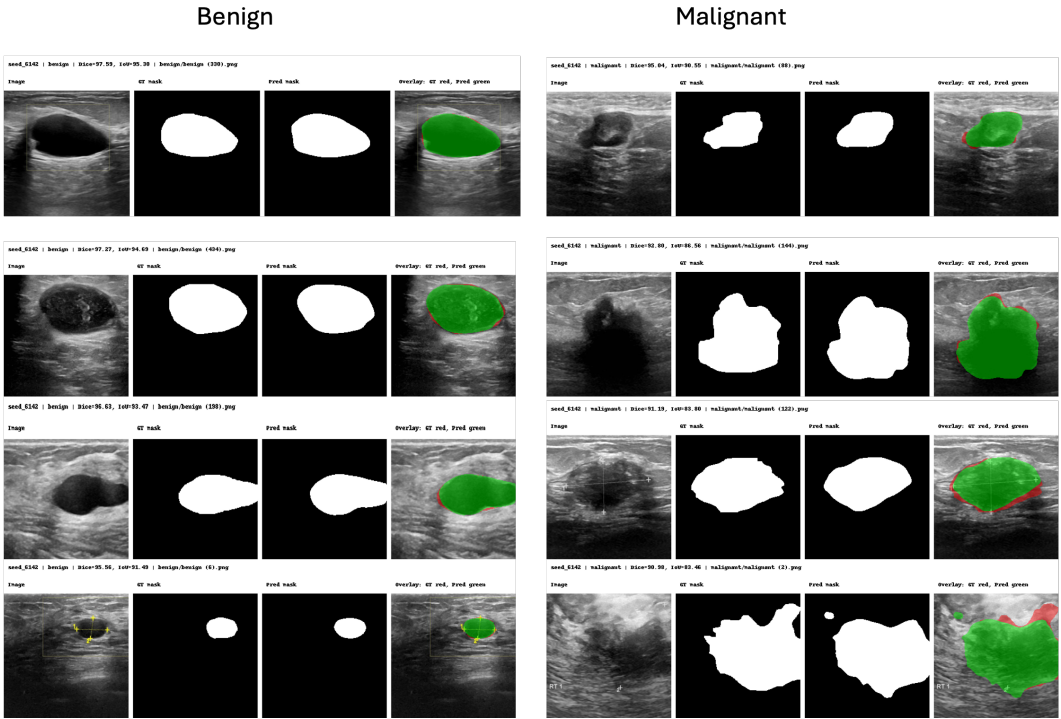


Figure 4. Qualitative Analysis of the lesion prediction masks for the downstream dense predictions on BUSI.

test splits. For Breast Cancer detection from ultrasound, we assess sensitivity (true positive rate/ recall) and specificity (true negative rate) using 5-fold cross validation. For the segmentation task, we use IoU (intersection-over-union) and dice coefficients.

5.2. Comparison with SOTA

Pneumonia Detection from Chest-XRay: PACL consistently outperforms the previous SOTA, GLCM-MAE (MICCAI 2025) [31] and PPAD (MICCAI 2024) [46], and achieves SOTA AUC with 2–2.3% improvement for pediatric pneumonia detection on the CXR dataset. Additionally, deterministic NEPA struggles severely with the high variance of CXR, while PACL tackles this with probabilistic bounds. We summarize the results in Tab. 1.

Breast Cancer Detection from Ultrasound: On the other

hand, PACL shows 2.1% improvement on sensitivity, and 1.7% improvement on AUROC of breast cancer detection on BUSI dataset over the previous SOTA, GLCM-MAE trained ViT, as shown in Tab. 2. Both CXR and BUSI classification improvements are statistically significant (CXR $p < 0.01$, BUSI $p < 0.03$).

Breast Cancer Segmentation in Ultrasound: Tab. 3 shows the segmentation performance for lesion segmentation on breast ultrasound. The downstream finetuning on the proposed PACL framework surpasses the SOTA IoU score reported by U-Next [50] by 1%. The robustness on the dense prediction task shows the suitability of the proposed framework for a broader category of downstream applications. For the segmentation task, overall the architectural components have approximately 89M parameters, and the pretraining plus finetuning takes roughly 48 GFLOPs. As

Variant	AUC	Acc.	F1	Loss	AUC	Acc.	F1
w/o PH	0.956	0.905	0.927	$\mathcal{L}_{cls}$	0.937	0.877	0.903
PH + Seq. CH	0.972	0.933	0.947	$\mathcal{L}_{cls} + \mathcal{L}_{uiip}$	0.962	0.902	0.926
PH + MTL CH	0.987	0.955	0.965	$\mathcal{L}_{cls} + \mathcal{L}_{uiip} + \mathcal{L}_{ocp}$	0.972	0.926	0.938
				$\mathcal{L}_{cls} + \mathcal{L}_{uiip} + \mathcal{L}_{ocp} + \mathcal{L}_{vs}$	0.987	0.955	0.965

Table 4. Ablation Study results on the Chest X-Ray dataset. (Left) Variation in architectural composition of PACL. **PH**: Probability Head, **CH**: Classification Head. (Right) Effect of different downstream loss components on PACL.

we have noted earlier, the ImageNet-1K pretraining is an one-time step, we do not actually gain any additional complexity during the pretrain-finetune phases for any downstream task on any medical dataset.

5.3. Qualitative Analysis

We visualize the predicted lesion segmentation masks in Fig. 4. Clearly, the proposed method shows effectiveness for both benign and malignant lesion segmentation.

5.4. Ablation Study

We report the ablations on the pneumonia classification task on the CXR dataset.

Architectural Composition: We show three variants in Tab. 4 (left) – (a) removing the probability head (uncertainty-guidance) and use only the ViT backbone with linear classifier – AUC drops by 3.1% ; (b) a cascaded pipeline where the output from the probabilistic head serves as the direct input for the final classification head (Sequential classification head); and (c) the proposed design where classification head and probabilistic head operate in a multi-task design, and the classification loss is explicitly regularized by uncertainty guidance.

Components of the Proposed Loss: We show in Tab. 4 (right) the effect of the base classification loss and the effect of the regularization components.

6. Conclusion

In this paper, we presented PACL, an autoregressive generative ViT tailored for robust medical image classification. PACL, is a pretraining framework that shifts medical image representation learning from deterministic point estimates to probabilistic distributional modeling. By predicting multivariate Gaussian distributions for causal anatomical patches, PACL intrinsically quantifies spatial aleatoric uncertainty, and prevents overconfidence during downstream adaptation via uncertainty-guided downstream loss. Extensive evaluations show that our proposed mechanism demonstrates SOTA resilience to confounding structures in X-rays (for classification) and noise in ultrasound (for both classification and segmentation). The robustness for both classification and segmentation shows the potential for broader applicability of PACL. In future, we will extend the framework

to 3D volumetric data, such as Coronary CT Angiography or Brain MRI.

References

- [1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 6
- [2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 5, 7
- [3] Mahmoud Assran and et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629, 2023. 1, 3
- [4] Shekoofeh Azizi et al. Big self-supervised models advance medical image classification. In *ICCV*, pages 3478–3488, 2021. 1, 2
- [5] Soumen Basu et al. Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In *MICCAI*, pages 423–433, 2022. 1, 2
- [6] Soumen Basu et al. Radformer: Transformers with global-local attention for interpretable and accurate gallbladder cancer detection. *Medical Image Analysis*, 83:102676, 2023. 6
- [7] Mark Chen et al. Generative pretraining from pixels. In *ICML*, pages 1691–1703. PMLR, 2020. 1, 3
- [8] Ting Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 1, 2
- [9] Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor. In *AAAI*, pages 383–392, 2022. 6
- [10] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, pages 8415–8424, 2021. 1
- [11] Thomas Defard et al. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, pages 475–489. Springer, 2021. 6
- [12] Alexey Dosovitskiy, Lucas Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [13] Florin C Ghesu et al. Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855, 2021. 1, 3
- [14] Rohit Girdhar et al. Omnimae: Single model masked pre-training on images and videos. In *CVPR*, pages 10406–10417, 2023. 1, 3, 6

- [15] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, pages 1705–1714, 2019. 6
- [16] Mayank Gupta, Soumen Basu, and Chetan Arora. How reliable are the metrics used for assessing reliability in medical imaging? In *MICCAI*, pages 149–158. Springer, 2023. 1, 3
- [17] Bonian Han, Yuktha Priya Masupalli, Xin Xing, and Gongbo Liang. Improving medical imaging model calibration through probabilistic embedding. In *BigData*, pages 4490–4496. IEEE, 2024. 1, 3
- [18] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 3
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 3, 6
- [21] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *EMNLP (Findings)*, pages 4246–4253, 2020. 3
- [22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 6
- [23] Shih-Cheng Huang et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digit. Med.*, 6(1), 2023. 1, 2
- [24] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 3
- [25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30, 2017. 3, 5
- [26] Daniel S Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 5
- [27] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 3, 6
- [28] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6
- [29] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021. 6
- [30] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024. 3
- [31] Chetan Madan et al. Focus on texture: rethinking pre-training in masked autoencoders for medical image classification. In *MICCAI*, pages 207–217, 2025. 3, 6, 7
- [32] Sergio Naval Marimont and Giacomo Tarroni. Implicit field learning for unsupervised anomaly detection in medical images. In *MICCAI*, pages 189–198. Springer, 2021. 6
- [33] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 7
- [34] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 6
- [35] Chantal Pellegrini et al. Xplainer: From x-ray observations to explainable zero-shot diagnosis. In *MICCAI*, pages 420–429, 2023. 6
- [36] Tal Reiss et al. Panda: Adapting pretrained features for anomaly detection and segmentation. In *CVPR*, pages 2806–2814, 2021. 6
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3
- [38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7
- [39] Mohammadreza Salehi et al. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, pages 14902–14912, 2021. 6
- [40] Abhishek Singh Sambyal, Narayanan C Krishnan, and Deepti R Bathula. Towards reducing aleatoric uncertainty for medical imaging tasks. In *ISBI*, pages 1–4. IEEE, 2022. 1, 3
- [41] Thomas Schlegl et al. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54, 2019. 6
- [42] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *ICLR*, 2022. 3, 4
- [43] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 3
- [44] Yashbir Singh et al. Deep learning-based uncertainty quantification for quality assurance in hepatobiliary imaging-based techniques. *Oncotarget*, 16:249, 2025. 1
- [45] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3
- [46] Zhichao Sun, Yuliang Gu, Yepeng Liu, Zerui Zhang, Zhou Zhao, and Yongchao Xu. Position-Guided Prompt Learning for Anomaly Detection in Chest X-Rays. In *MICCAI*, 2024. 3, 6, 7

- [47] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 3
- [48] Ekin Tiu et al. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng.*, 6(12):1399–1406, 2022. 3, 6
- [49] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, pages 32–42, 2021. 3
- [50] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, pages 23–33. Springer, 2022. 7
- [51] Tiange Xiang et al. Squid: Deep feature in-painting for unsupervised anomaly detection. In *CVPR*, pages 23890–23901, 2023. 6
- [52] Sihan Xu et al. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025. 1, 3, 6
- [53] He Zhao et al. Anomaly detection for medical images using self-supervised and translation-consistent features. *IEEE TMI*, 40(12):3641–3651, 2021. 6
- [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 6
- [55] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018. 7
- [56] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 2