Anatomy-Anchored Self-Supervision: Distilling Vision Foundation Models for Invariant Ultrasound Representation

Chunzheng Zhu[1], Yijun Wang[1], Jianxin Lin[1]*, Feng Wang[1], Hongwei Wang[1], Lei Zhao[1], Shengli Li[2], and Kenli Li[1]

[1] Hunan University, Changsha, China
linjianxin@hnu.edu.cn
[2] Shenzhen Maternity and Child Healthcare Hospital, Shenzhen, China

Abstract. Self-supervised pre-training paradigm has gained increasing prominence for learning transferable representations in medical imaging, yet existing methods for ultrasound (US) images operate at the image or frame level, overlooking the anatomical context for clinical-aligned representation learning. In this work, we propose an Anatomy-Anchored UltraSound Self-Supervision framework (AnaUS) that shifts representation learning from generic visual regions to clinically meaningful anatomical structures. Utilizing a learnable latent prompt engine alongside a one-time domain adaptation on existing public image–mask pairs, we empower the LP-SAM module to achieve annotation-free anatomy delineation at scale. Building upon this anatomical grounding, we propose a dual-policy self-supervised learning paradigm consisting of inter-view semantics-aware anatomy-separating alignment and contextual core-region prediction to enhance representation learning. Specifically, the former enforces feature invariance within identical anatomical regions while promoting discriminability across distinct structures; the latter compels the model to reconstruct corrupted regions, thereby capturing fine-grained structural details. Extensive evaluations on six public datasets demonstrate that AnaUS consistently outstrips current state-of-the-art methods while maintaining the computational efficiency essential for clinical deployment. Code is available at https://github.com/zhcz328/ANAUS.

Keywords: Self-supervised Learning · Ultrasound Imaging · Anatomy-level Contrast · Foundation Model Adaptation.

1 Introduction

Ultrasound (US) imaging is a cornerstone of clinical diagnostics owing to its real-time capability, portability, and non-ionizing nature [4,33]. Deep neural networks have demonstrated considerable promise in automating US image analysis [11,8,18], yet their data-hungry training regimes remain at odds with the

* Corresponding author.

limited availability of annotated clinical data. Self-supervised learning (SSL) addresses this bottleneck by deriving supervision from the data itself, enabling pre-training on large unlabeled corpora followed by efficient fine-tuning [16,6,14].

Current SSL strategies for ultrasound broadly fall into two categories. *Frame-level* approaches leverage temporal coherence in US video to construct positive pairs [8,2,26], yet they treat each frame holistically and risk encoding background artifacts and speckle noise rather than diagnostically relevant anatomy. *Region-level* methods borrowed from natural images [17,28] attempt finer-grained learning, but their reliance on generic segmentation (*e.g.*, K-means clustering or image-computable proposals) fails to isolate meaningful anatomical structures in US images, where boundaries are characteristically low-contrast and speckle-corrupted [3]. Neither paradigm aligns with clinical reasoning, which is inherently *anatomy-centric*: diagnostic interpretation focuses on the morphology, texture, and spatial relationships of specific anatomical landmarks.

To ensure congruence with clinical reasoning, the US representation learning paradigm must prioritize anatomical structures over stochastic image patches as fundamental contrastive units. *Attaining such structural awareness requires reconciling the inherent domain disparity between natural and medical imagery.* To this end, our proposed AnaUS performs self-supervised pre-training directly at the anatomy level. This approach is powered by LP-SAM, a domain-specialized module that utilizes a learnable latent prompt engine and lightweight fine-tuning to bridge the domain gap, thereby achieving automated anatomy discovery.

Utilizing this delineation capability, AnaUS integrates two synergistic learning mechanisms. The first, view-invariant anatomical contrastive matching, regularizes the representation space to ensure invariance for the same structure under diverse views while maximizing separation between distinct anatomical entities. Such a formulation introduces explicit inter-anatomy discriminability, a property typically absent in generic contrastive frameworks. Complementing this, an anatomical contextual prediction task targets the reconstruction of corrupted core regions, compelling the encoder to internalize both local textural intricacies and global structural coherence. These dual objectives operate across a multi-scale feature hierarchy, ensuring that the learned representations encapsulate anatomical traits at hierarchical granularities.

Comprehensive evaluations across six public ultrasound benchmarks, encompassing lung, breast, thyroid, and cardiac imaging, validate the multi-task efficacy of AnaUS in various downstream clinical scenarios. The proposed framework sets a new state-of-the-art, consistently surpassing both general-purpose SSL methods and previous ultrasound-specific pre-training approaches. As a result, our work underscores the pivotal role of anatomically meaningful priors in guiding the representation learning process for medical imagery.

Our contributions can be summarized as: ***(i)*** We propose AnaUS, the first anatomy-level self-supervised pre-training framework designed explicitly for ultrasound imaging, shifting the contrastive unit from generic regions to clinically relevant structures. ***(ii)*** We introduce LP-SAM with a learnable latent prompt engine that empowers fully automated, expert-free structural discovery in ultra-

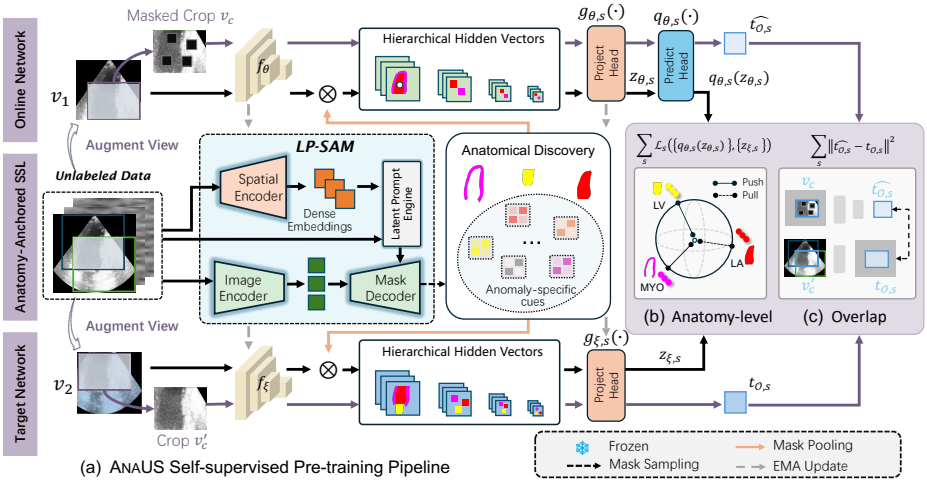

Fig. 1: Overview of the ANAUS framework. *(i)* LP-SAM provides anatomical masks $\mathcal{M}(x)=\{m_k\}_{k=1}^{K}$ for unlabeled US images via a learnable latent prompt engine (Sec. 2.1). *(ii)* The encoder f_θ is jointly optimized by an anatomy-level contrastive learning objective across augmented view pairs $(\mathbf{v}_1, \mathbf{v}_2)$ (Sec. 2.2) and a contextual prediction objective over corrupted core regions (Sec. 2.3).

sound imagery through targeted latent-space adaptation. ***(iii)*** We demonstrate consistent state-of-the-art performance across diverse US datasets and tasks, establishing a new benchmark for ultrasound self-supervised pre-training.

2 Method

Let $\mathcal{X}=\{x_i\}_{i=1}^{N}$ be a corpus of unlabeled US images. ANAUS (Fig. 1) seeks an encoder $f_\theta: \mathcal{X}\to\mathcal{F}$ whose representations are invariant within anatomical structures yet discriminative across them, without any quality or segmentation annotations. Given x, LP-SAM produces a per-image mask set $\mathcal{M}(x)=\{m_k\}_{k=1}^{K}$, $m_k \in \{0,1\}^{H\times W}$. Two augmented views $\mathbf{v}_1=t(x)$ and $\mathbf{v}_2=t'(x)$, $t, t'\sim\mathcal{T}$, are fed to an asymmetric online–target encoder pair optimized jointly for representation.

2.1 Latent-Prompted Domain Adaptation of LP-SAM

SAM [19] exhibits zero-shot segmentation capability on natural images, yet its direct application to US data yields suboptimal masks owing to the domain gap. LP-SAM (Fig. 2) resolves this via a *two-stage lightweight adaptation* that preserves SAM's capacity while injecting domain-specific inductive biases.

Stage 1: Adapter-based domain specialization. Following [29], we insert learnable adapters into each ViT block of SAM's image encoder while keeping the pre-trained weights frozen. Each adapter implements a bottleneck projection

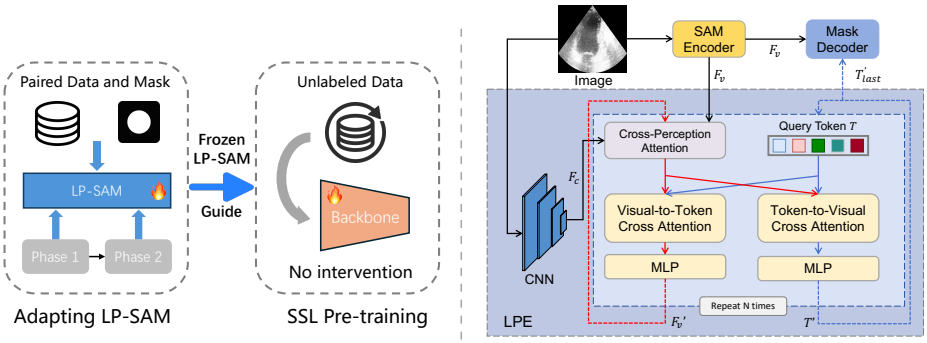

Fig. 2: *(i) Left*: The overall training pipeline of AnaUS. Phase 1 fine-tunes LP-SAM adapters on paired US image–mask data; Phase 2 trains the latent prompt engine (LPE) with LP-SAM frozen. The resulting LP-SAM is then frozen to guide SSL pre-training on unlabeled data. *(ii) Right*: The LP-SAM architecture. In LPE, Cross-Perception Attention (CPA) fuses global ViT and local CNN features; bidirectional cross-attention iteratively refines query tokens into latent prompt embedding T'_{last} for annotation-free mask generation.

$\phi: \mathbb{R}^d \to \mathbb{R}^{d'}$ $(d' \ll d)$ with a residual bypass, trained on existing US image–mask pairs [23,30,20] via $\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{Dice}}$. The Dice term proves critical for small-structure US targets with severe foreground–background imbalance.

Stage 2: Latent prompt engine. Annotation-free deployment demands prompt-free mask generation. Let $F_v \in \mathbb{R}^{c\times h\times w}$ and $F_c \in \mathbb{R}^{c\times h\times w}$ denote the global ViT feature and local CNN feature, respectively. A Cross-Perception Attention (CPA) module fuses complementary representations as:

$$F_{vc} = LN(MHCA(F_v, F_c, F_c) + F_v)\,, \tag{1}$$

where MHCA$(\cdot)$ is multi-head cross-attention [27] with F_v as query and F_c as key and value. A learnable query token $T \in \mathbb{R}^{k\times c}$ is then iteratively refined over N rounds of bidirectional cross-attention:

$$F'_v = MLP(LN(MHCA(F_{vc}, T, T) + F_{vc}))\,, \tag{2}$$
$$T' = MLP(LN(MHCA(T, F_{vc}, F_{vc}) + T))\,, \tag{3}$$

where each round feeds the updated (F'_v, T') and original F_c as new inputs. The final token T'_{last} serves as the task-conditioned prompt embedding for SAM's mask decoder, enabling fully annotation-free anatomical mask generation. During Phase 2, SAM and its adapters are frozen; both stages are *one-time* procedures that do not participate in subsequent SSL pre-training.

2.2 View-invariant Anatomical Contrastive Matching

Previous SSL methods treat entire images or random crops as the unit of comparison, mixing diagnostically relevant anatomy with noisy background. AnaUS

operates at the level of anatomical structures identified by LP-SAM, ensuring that representation learning is grounded in clinically meaningful semantics.

View construction and mask alignment. Given an image x, two views $\mathbf{v}_1 = t(x)$ and $\mathbf{v}_2 = t'(x)$ are generated by sampling augmentations $t \sim \mathcal{T}$, $t' \sim \mathcal{T}'$. The corresponding anatomical masks are transformed identically, yielding aligned mask sets m and m'. From each set, K masks are sampled to form a balanced batch of anatomical instances.

Multi-scale mask pooling. Each view is encoded by a dual-branch shared backbone f_θ, producing feature maps $\{f_s\}$ at S spatial scales. For each mask m_s downsampled to scale s, a mask-pooled hidden vector is computed as:

$$h_s = \frac{\sum_{i,j} m_s[i,j] \cdot f_s[i,j]}{\sum_{i,j} m_s[i,j]}, \tag{4}$$

computing the area-normalized mean feature within the region of interest, eliminating background contamination. This approach is more targeted than the heuristic-driven region proposals used in prior object-centric methods [17,28], which are ill-suited to the low-contrast boundaries common in US images.

Contrastive objective. We adopt an asymmetric online–target architecture following BYOL [14]. The online network maps h_s to a projection $z_{\theta,s} = g_{\theta,s}(h_s)$ and subsequently to a prediction $q_{\theta,s}(z_{\theta,s})$; the target network produces $z'_{\xi,s} = g_{\xi,s}(h'_s)$ with parameters ξ updated via exponential moving average. After ℓ_2-normalization $\bar{q}_{\theta,s}{=}q_{\theta,s}/\|q_{\theta,s}\|_2$ and $\bar{z}'_{\xi,s}{=}z'_{\xi,s}/\|z'_{\xi,s}\|_2$, the separating loss encourages matching anatomy across views while repelling distinct structures:

$$\ell^{(s)}(\theta;\xi) = -\log \frac{\exp\big(\langle \bar{q}_{\theta,s},\, \bar{z}'_{\xi,s}\rangle/\tau\big)}{\exp\Big(\langle \bar{q}_{\theta,s},\, \bar{z}'_{\xi,s}\rangle/\tau\Big) + \sum_{n\neq \mathrm{pos}} \exp\big(\langle \bar{q}_{\theta,s},\, \bar{z}'_{n,s}\rangle/\tau\big)}, \tag{5}$$

where τ is the temperature and the summation over n ranges over all non-matching anatomy instances in the batch. The total anatomy-level contrastive loss $\mathcal{L}_{\mathrm{ana}}(\theta;\xi) = \frac{1}{K}\sum_{k=1}^{K}\sum_{s=1}^{S}\Big(\ell^{(s)}(\theta;\xi) + \tilde{\ell}^{(s)}(\theta;\xi)\Big)$ symmetrises views and aggregates across scales, where $\tilde{\ell}^{(s)}$ denotes the symmetric counterpart obtained by swapping the roles of views. Operating across S scales allows the model to capture anatomy at both coarse structural and fine textural granularity.

2.3 Contextual Prediction over Anatomical Regions

Complementing $\mathcal{L}_{\mathrm{ana}}$ with a reconstruction objective over corrupted anatomical regions forces f_θ to capture structural context beyond surface-level appearance [24,15,31]. Unlike random masking [15], our approach targets the *core anatomical region*, *i.e.*, the overlapping crop between $\mathbf{v}_1$ and $\mathbf{v}_2$, ensuring recovery of clinically meaningful content rather than arbitrary background.

Specifically, we extract the overlapping crop shared by $\mathbf{v}_1$ and $\mathbf{v}_2$, partition it into patches, and apply a damage function $D(I;\alpha,\beta,\gamma)$ that combines three

Table 1: **Performance comparison.** Best in **bold**, second-best underlined. †: LP-SAM masks; ++: expanded pre-training on `BF+CM` dataset. POCUS and BUSI are classification benchmarks (accuracy and F1, %); UDIAT-B, TN3K, BUSI, and HMC-QU are segmentation benchmarks (PPV/Sensitivity/Dice in %, HD). All differences vs. AnaUS on UDIAT-B are significant at $p<0.01$.

			POCUS				BUSI		UDIAT-B			TN3K		BUSI		HMC-QU	
	Method	Mask	COVID	Pneu.	Reg.	Acc↑	Acc↑	F1↑	PPV↑	Sens.↑	Dice↑	Dice↑	HD↓	Dice↑	HD↓	Dice↑	HD↓
General	Supervised	—	83.7	82.1	86.5	85.0	71.3	82.8	81.5	74.7	77.9	81.57	34.31	80.55	32.91	80.70	20.72
	ImageNet [9]	—	79.5	78.6	88.6	84.2	84.9	81.8	82.0	80.8	81.4	81.72	33.78	81.03	32.64	81.50	21.08
	SimCLR [6]	—	83.2	89.4	87.1	86.4	74.6	86.3	77.4	80.8	79.1	82.10	33.25	81.12	32.45	82.14	20.49
	MoCo v2 [7]	—	79.7	81.4	88.9	84.8	77.8	82.8	72.6	78.6	75.4	82.17	33.54	80.13	32.59	81.42	19.96
Object-Centric	SlotCon [28]	SG	82.2	80.9	87.3	82.8	74.1	80.1	85.3	79.9	82.7	82.02	32.15	80.28	31.74	81.51	21.23
	SlotCon†	LP-SAM	91.8	88.3	90.1	89.2	78.9	89.5	88.1	81.4	84.0	84.12	28.67	82.50	31.55	80.11	21.05
	DetCon [17]	MCG	81.9	80.4	87.1	82.4	70.1	79.5	84.0	79.3	81.1	81.85	32.40	80.12	32.90	80.07	21.11
	DetCon†	LP-SAM	86.2	89.4	91.0	88.1	82.3	91.9	87.2	81.7	83.4	83.16	31.75	82.42	31.80	82.05	20.10
	R2O [12]	K-means	82.3	81.7	88.3	84.5	72.8	83.5	82.7	80.5	81.6	81.65	32.62	81.78	31.68	81.16	21.10
	R2O†	LP-SAM	89.5	91.2	91.8	90.9	79.3	88.9	83.6	81.2	82.2	83.70	29.27	82.24	31.42	82.21	20.82
US-Specific	USCL++ [8]	—	86.3	90.4	93.4	90.9	85.6	93.2	88.2	80.5	84.4	84.01	29.25	82.12	31.45	82.64	20.49
	USF-MAE++ [22]	—	90.1	92.8	94.0	93.2	85.6	93.2	88.6	81.1	84.5	84.20	29.70	82.30	31.35	82.48	20.22
	USUCL++ [2]	—	89.5	94.9	93.3	92.3	84.2	93.0	88.4	81.2	84.1	83.50	28.98	82.35	31.60	82.05	20.35
	AWCL++ [10]	—	88.3	87.6	91.5	91.7	84.5	91.2	87.4	80.9	83.8	83.42	30.15	81.78	31.90	81.95	20.65
	IVPP++ [26]	—	91.5	95.1	93.2	93.2	83.5	92.5	89.1	80.3	83.6	83.85	28.66	81.85	31.06	82.65	19.88
	AnaUS (ours)	LP-SAM	**93.4**	**95.4**	**96.1**	**96.2**	**87.1**	**94.3**	**92.7**	**82.3**	**85.9**	**85.42**	**26.31**	**83.35**	**30.45**	**83.48**	**19.41**

perturbation types tailored to US artefacts:

$$D(x;\,\alpha,\beta,\gamma) = \big(\mathcal{O}(\cdot,\gamma)\circ\mathcal{N}(\cdot,\beta)\circ\mathrm{Shuff}(\cdot,\alpha)\big)(x), \tag{6}$$

where $\mathrm{Shuff}(\cdot,\alpha)$ performs local pixel shuffling with probability α to simulate tissue displacement under probe pressure, $\mathcal{N}(0,\beta)$ injects Gaussian noise mimicking speckle interference, and $\mathcal{O}(\gamma)$ applies random occlusion to emulate acoustic shadowing. Finally, the f_θ processes the damaged crop to obtain features $\hat{\mathbf{t}}_{o,s}$, while the f_ξ processes the intact crop to yield $\mathbf{t}_{o,s}$. The loss is defined as: $\mathcal{L}_{\mathrm{ctx}}(\theta;\xi) = \sum_{s=1}^{S} \|\hat{\mathbf{t}}_{o,s} - \mathbf{t}_{o,s}\|_2^2$. The pre-training loss integrates both objectives:

$$\mathcal{L}_{\mathrm{total}}(\theta;\xi) = \mathcal{L}_{\mathrm{ana}}(\theta;\xi) + \lambda_{\mathrm{ctx}}\cdot\mathcal{L}_{\mathrm{ctx}}(\theta;\xi), \tag{7}$$

where λ_{ctx} controls the relative contribution of contextual prediction. After pre-training, only the online encoder f_θ is retained for downstream evaluation; all projection, prediction, and target components are discarded.

3 Experiments

3.1 Experimental Setup

Implementation Details. AnaUS is implemented in PyTorch with a ResNet-18 FPN backbone [21] initialized from ImageNet weights [9]. The training consists of two phases on two RTX 4090 GPUs. First, LP-SAM adapter fine-tuning runs for 200 epochs (Adam, lr $= 5\times10^{-4}$, batch 16), followed by 100 epochs of latent prompt engine training with SAM frozen. Second, SSL pre-training proceeds for

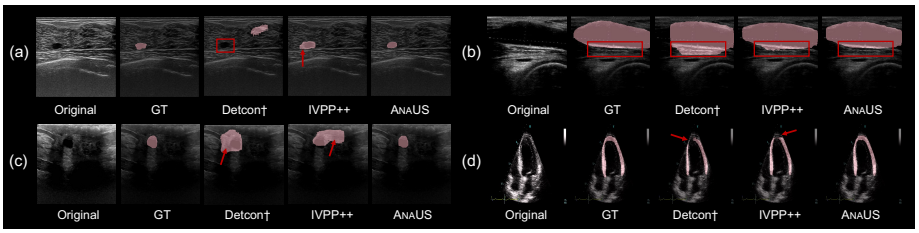

Fig. 3: Qualitative downstream segmentation comparison on UDIAT-B, TN3K, BUSI, and HMC-QU. Left to right: Input, ground truth (GT), baselines, AnaUS.

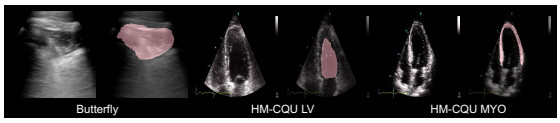

Fig. 4: Zero-shot LP-SAM generalization of out-of-distribution structures (bat sign, LA, MYO). The latent T'_{last} enables domain-agnostic anatomical discovery.

300 epochs via LARS with a cosine-decay schedule and three-epoch warm-up (base lr $= 0.3$, batch 192). Key hyperparameters are set as follows: $\tau = 0.1$, $K = 16$, $S = 4$, $\lambda_{\text{ctx}} = 1$, $N = 4$, and damage parameters $(\alpha, \beta, \gamma) = (0.15, 0.2, 0.05)$.

Datasets, Baselines, and Evaluation Protocol. We utilize distinct datasets across different stages. For LP-SAM fine-tuning, we use DDTI [23] (637 images), TG3K [30] (3,585 images), and CAMUS [20] (57,696 images). SSL pre-training is conducted on a combined dataset (denoted `BF+CM`) comprising Butterfly [5] and CAMUS [20]. Downstream evaluation spans **classification** on POCUS [4] and BUSI [1] (evaluated via accuracy and F1), and **segmentation** on UDIAT-B [32], TN3K [13], BUSI, and HMC-QU (PPV, Sensitivity, Dice, and Hausdorff distance). All downstream tasks employ five-fold cross-validation. We benchmark against three categories: *(i)* **general SSL** (ImageNet [9], SimCLR [6], FixMatch [25], MoCo v2 [7]); *(ii)* **object-centric methods** (SlotCon [28], DetCon [17], R2O [12]), with † denoting LP-SAM mask substitution; and *(iii)* **US-specific pre-training methods** (USCL [8], USF-MAE [22], USUCL [2], AWCL [10], IVPP [26]), with ++ denoting models pre-trained on the `BF+CM` set.

3.2 Comparison with State-of-the-Art Methods

As shown in Table 1, AnaUS achieves the best performance across benchmarks.

Classification. AnaUS attains **96.2**% accuracy on POCUS and **87.1**% on BUSI, surpassing the strongest US-specific baseline IVPP++ by +**3.0**% and +**3.6**%, respectively. Per-class POCUS margins over the supervised baseline reach +9.7% (COVID-19), +13.3% (Pneumonia), and +9.6% (Regular), highlighting the adaptability of anatomy-level SSL to pathology-discriminative tasks.

Table 2: Ablation analysis of AnaUS on UDIAT-B (PPV/Sensitivity/Dice, %). *Left*: component analysis ($\tilde{\ell}$: symmetric loss, LPE: latent prompt engine, MG: multi-scale contrast, CP: contextual prediction). *Right*: CEP function decomposition. Statistical significance vs. full model (paired t-test): † $p<0.05$; ‡ $p<0.01$.

$\tilde{\ell}$	LPE	MG	CP	PPV	Sens.	Dice
✓			✓	77.0†	72.3‡	74.3†
	✓	✓		88.5†	78.5†	81.2†
	✓	✓	✓	84.4‡	74.3‡	77.6†
✓		✓	✓	72.4†	68.1†	70.2‡
✓	✓		✓	89.5†	80.5†	83.8†
✓	✓	✓		91.1†	81.6†	84.9†
✓	✓	✓	✓	**92.7**	**82.3**	**85.9**

Damage Component	PPV	Sens.	Dice
AnaUS (Full)	**92.7**	**82.3**	**85.9**
w/o Shuff$(\cdot, \alpha)$	91.8†	82.0†	84.9†
w/o $\mathcal{N}(0, \beta)$	92.3†	81.9†	85.4†
w/o $\mathcal{O}(\gamma)$	91.9†	82.2†	85.1†
w/o CEP entirely	91.1†	81.6†	84.7†

Segmentation. On UDIAT-B, AnaUS reaches **85.9**% Dice with **92.7**% PPV, with consistent HD reductions across TN3K, BUSI, and HMC-QU. A critical observation is the ***mask-source effect***: replacing generic mask sources (multi-scale combinatorial grouping/MCG, semantic grouping/SG, and K-means) with LP-SAM substantially elevates object-centric baselines across all benchmarks (+2.3% Dice for DetCon, +1.3% for SlotCon on UDIAT-B), confirming that ***anatomically precise masks are the decisive factor*** for US pre-training quality rather than the contrastive objective formulation alone.

Visual Results. Fig. 3 qualitatively corroborates AnaUS's superior boundary precision. While DetCon† shows coarse localization and IVPP++ over-smooths challenging structures (*e.g.*, thyroid nodules in TN3K and myocardial edges in HMC-QU), AnaUS maintains sharp delineation. Furthermore, Fig. 4 confirms LP-SAM's zero-shot generalization to unseen anatomy, demonstrating that iterative cross-attention (Sec. 2.1) produces structurally generalizable prompt embeddings T'_{last} rather than dataset-specific pattern matching.

3.3 Comprehensive Ablation Analysis

Component Analysis. Table 2 (left) dissects each design choice. Substituting LP-SAM with vanilla SAM (*i.e.*, w/o LPE) incurs the sharpest degradation (−15.7% Dice), underscoring ***domain-adapted mask*** as a prerequisite for grounded SSL. Removing symmetric contrast ($\tilde{\ell}$) drops Dice by 8.3% ($p<0.01$), while disabling multi-scale contrast (MG) and contextual prediction (CP) further reduces Dice by 2.1% and 1.0% respectively ($p<0.05$), with each component proving complementary.

Damage Function Decomposition. Table 2 (right) isolates each perturbation in the CEP damage function. Pixel shuffling yields the largest gain, followed by occlusion and Gaussian noise, with each component simulating distinct US imaging artefacts; removing all three degrades performance by 1.2%.

4 Conclusion

We present AnaUS, an anatomy-anchored self-supervised pre-training framework that shifts the contrastive unit from arbitrary patches to clinically meaningful structures. LP-SAM bridges the natural-to-ultrasound domain gap via lightweight adapters and a learnable latent prompt engine, enabling annotation-free anatomical discovery. Inter-view anatomy-level contrast and contextual prediction jointly enforce inter-anatomy discriminability and structural coherence. Extensive evaluations confirm that AnaUS establishes a principled and clinically aligned foundation for future research in medical self-supervised learning.

Acknowledgments. This research was partially supported by the National Major Scientific Instrument Development Project of the National Natural Science Foundation of China (Grant No. 62227808), the National Natural Science Foundation of China (Grants No. 62472157), and the Hunan Provincial Graduate Research Innovation Project (Grant No. CX20250587).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 423–433. Springer (2022)
3. Biradar, N., Dewal, M.L., Rohit, M.K.: Speckle noise reduction in b-mode echocardiographic images: A comparison. IETE Technical Review **32**(6), 435–453 (2015)
4. Born, J., Wiedemann, N., Cossio, M., Buhre, C., Brändle, G., Leidermann, K., Goulet, J., Aujayeb, A., Moor, M., Rieck, B., et al.: Accelerating detection of lung pathologies with explainable ultrasound image analysis. Applied Sciences **11**(2), 672 (2021)
5. ButterflyNetwork: Butterfly videos. https://www.butterflynetwork.com/index.html (2020), accessed: September 20, 2020
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
8. Chen, Y., Zhang, C., Liu, L., Feng, C., Dong, C., Luo, Y., Wan, X.: Uscl: pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th International conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 627–637. Springer (2021)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

10. Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Anatomy-aware contrastive representation learning for fetal ultrasound. In: European Conference on Computer Vision. pp. 422–436. Springer (2022)
11. Gao, Y., Maraci, M.A., Noble, J.A.: Describing ultrasound video content using deep convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). pp. 787–790. IEEE (2016)
12. Gokul, A., Kallidromitis, K., Li, S., Kato, Y., Kozuka, K., Darrell, T., Reed, C.J.: Refine and represent: Region-to-object representation learning. arXiv preprint arXiv:2208.11821 (2022)
13. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in biology and medicine **155**, 106389 (2023)
14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
17. Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10086–10096 (2021)
18. Jiang, C., Zhu, C., Guo, H., Tan, G., Liu, C., Li, K.: Mcbl-unet: A hybrid mamba-cnn boundary enhanced light-weight unet for placenta ultrasound image segmentation. IEEE Journal of Biomedical and Health Informatics (2025)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
20. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
22. Megahed, Y., Ducharme, R., Erman, A., Walker, M., Hawken, S., Chan, A.D.: Usf-mae: Ultrasound self-supervised foundation model with masked autoencoding. arXiv preprint arXiv:2510.22990 (2025)
23. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International symposium on medical information processing and analysis. vol. 9287, pp. 188–193. SPIE (2015)
24. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience **2**(1), 79–87 (1999)
25. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with

consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
26. VanBerlo, B., Wong, A., Hoey, J., Arntfield, R.: Intra-video positive pairs in self-supervised learning for ultrasound. arXiv preprint arXiv:2403.07715 (2024)
27. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
28. Wen, X., Zhao, B., Zheng, A., Zhang, X., Qi, X.: Self-supervised visual representation learning with semantic grouping. Advances in neural information processing systems **35**, 16423–16438 (2022)
29. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
30. Wunderling, T., Golla, B., Poudel, P., Arens, C., Friebe, M., Hansen, C.: Comparison of thyroid segmentation techniques for 3d ultrasound. In: Medical Imaging 2017: Image Processing. vol. 10133, pp. 346–352. SPIE (2017)
31. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
32. Yap, M.H., Goyal, M., Osman, F., Martí, R., Denton, E., Juette, A., Zwiggelaar, R.: Breast ultrasound region of interest detection and lesion localisation. Artificial Intelligence in Medicine **107**, 101880 (2020)
33. Zhu, C., Lin, J., Tan, G., Zhu, N., Li, K., Wang, C., Li, S.: Advancing ultrasound medical continuous learning with task-specific generalization and adaptability. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3019–3025. IEEE (2024)