CPR: Chained Perceptual Refinement for Coarse-to-Fine Medical Image Classification

Si-Yuan Lu[1,*], Hanruo Zhu[2,*], Ziquan Zhu[2,*], Gaojie Jin[3], Zeyu Fu[3], Lu Yin[4](✉), Ke Li[3], Lu Liu[3], and Tianjin Huang[3](✉)

[1] Nanjing University of Posts and Telecommunications, Nanjing, China
[2] University of Leicester, Leicester, UK
[3] University of Exeter, Exeter, UK
[4] University of Surrey, Guildford, UK
t.huang2@exeter.ac.uk;l.yin@surrey.ac.uk
*Equal contribution.

Abstract. High-resolution medical images contain fine-grained, spatially sparse cues that are critical for diagnosis, yet preserving full resolution incurs substantial computational and memory costs. Most deep models process images uniformly, leading to redundant computation or loss of diagnostic detail under downsampling. We propose Chained Perceptual Refinement (CPR), a coarse-to-fine framework that formulates medical image analysis as a sequential global-to-local decision process. Starting from a low-resolution global view, CPR dynamically predicts the location and spatial extent of refinement regions, extracts high-resolution evidence from the original image, and incrementally integrates it with global context. By keeping the backbone input size fixed while contracting the perceptual field, CPR preserves diagnostic fidelity with constant peak GPU memory. Extensive experiments on five medical imaging datasets and multiple backbone architectures demonstrate that CPR consistently outperforms both fixed-resolution and multi-scale state-of-the-art baselines, achieving improvements of up to 2.27% over the second-best method. It also achieves up to a 19.6× reduction in GFLOPs at matched accuracy, establishing a superior accuracy–efficiency trade-off for high-resolution medical image analysis. The code is available at: GitHub.

Keywords: Medical Image Classification · Chained Perceptual Refinement · Coarse-to-Fine Visual Processing · Adaptive Computation

1 Introduction

High-resolution medical images are central to modern clinical workflows, as diagnostically decisive cues are often fine-grained and spatially sparse. Across modalities such as digital pathology, high-resolution radiography, and CT, clinically meaningful evidence may manifest as subtle textural variations, small morphological structures, or tiny lesions occupying only limited regions[27,25]. Consequently, many medical tasks including detection, segmentation, and treatment planning, must jointly leverage global anatomical context and local high-

frequency details [20,7,2]. *This creates a fundamental trade-off*: preserving diagnostic resolution incurs substantial memory and computation. As shown in Fig. 1(a), increasing input resolution generally improves accuracy, yet GPU memory consumption rises sharply and quickly becomes impractical. Conversely, aggressive downsampling reduces cost at the expense of diagnostic fidelity (Fig. 1(b)) , exposing the limitation of fixed-resolution paradigms that uniformly allocate computation across spatially heterogeneous data.

While dominant deep learning architectures have achieved remarkable success, they predominantly rely on spatially uniform computation [16]. Standard approaches process inputs at fixed dimensions or utilize dense patch-based extraction [24,21], uniformly allocating compute across an image regardless of regional informativeness [8,4]. Multi-scale architectures (e.g., pyramidal features) incorporate multiple resolutions, yet their computation remains predefined and often redundant [12,19,17,14]. More importantly, these paradigms are passive: models process inputs with fixed resolution and static allocation, without adaptively deciding *where* and *at what scale* to focus. *In contrast*, human perception is sequential and selective. Clinicians rarely analyze medical images in a single pass; instead, they form a global impression and progressively refine attention toward diagnostically relevant regions, integrating global context with localized evidence. While recent advancements such as AdaptiveNN [26,13] explore active, human-inspired vision mechanisms by explicitly modeling sequential fixations, they predominantly rely on extracting predefined, fixed-size crops. In medical image analysis, where subtle diagnostic evidence is strictly anchored to its broader anatomical context, such unconstrained, disjointed patch sampling risks redundant observations and disrupts the spatial hierarchy.

Motivated by foveal vision, we introduce Chained Perceptual Refinement (`CPR`), a perception-inspired framework that formulates medical image analysis as a dynamic, coarse-to-fine sequential decision process. Starting from a low-resolution global view, `CPR` iteratively allocates a localized *perceptual field* by dynamically predicting refinement parameters, specifically, both top-left coordinates and spatial extent, at each step. The selected regions are extracted from the original high-resolution image and incrementally fused with the global context to update the model's prediction. Unlike Adaptive NN, which mainly uses fixed-size local observations, `CPR` jointly adapts crop location and scale; unlike OverLoCK, it explicitly revisits the original high-resolution image rather than relying on feature-level context mixing; and unlike PAMIL, it targets image-level classification instead of WSI-level MIL. We summarize our main contributions as follows:

- ⋆ **Method.** We propose `CPR`, a framework that transitions medical image analysis from static processing to an active, sequential paradigm. By explicitly modeling global-to-local transitions through dynamic scale contraction, `CPR` progressively restricts its perceptual field to diagnostically relevant regions, emulating the clinical coarse-to-fine diagnostic workflow.
- ⋆ **SOTA Performance.** `CPR` demonstrates superior performance across four modalities and multiple architectures. Specifically, it consistently surpasses

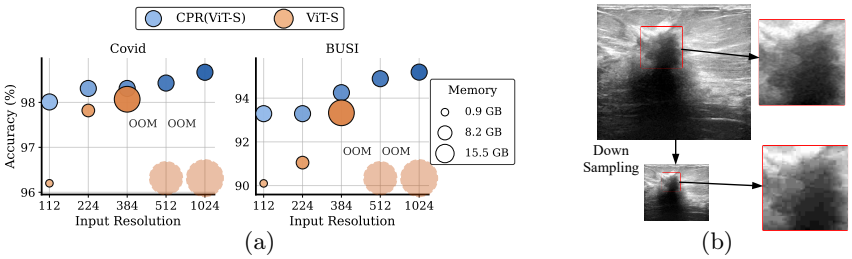

Fig. 1. Resolution–efficiency trade-off in medical image analysis. (a) Accuracy and peak GPU memory at different input resolutions on COVID and BUSI (RTX 3090, batch size 64). Bubble size denotes memory; OOM indicates infeasible settings. (b) Downsampling removes fine-grained details.

the current state-of-the-art (SOTA) baseline, Adaptive NN [26], across the Shenzhen, Covid, BUSI, BreaKHis, and EyePACs datasets, with improvements of 2.27%, 0.24%, 1.92%, 0.45%, and 1.42%, respectively.

⋆ **Memory and Computational Efficiency.** By employing a selective refinement strategy with a fixed backbone input size, `CPR` ensures constant peak GPU memory regardless of original image resolution, effectively bypassing quadratic memory scaling. This framework establishes a superior accuracy–efficiency trade-off, delivering up to a 19.6× reduction in GFLOPs compared to dense high-resolution processing at equivalent accuracy levels.

2 Chained Perceptual Refinement Framework

We propose Chained Perceptual Refinement (`CPR`), a sequential coarse-to-fine framework that dynamically allocates perceptual fields while maintaining constant computational cost (Fig. 2). The framework consists of four core modules: a *Shared Encoder*, a *Perceptual State* update mechanism, a *Policy Network* for refinement, and a *Prediction Head* for task supervision. Instead of uniformly processing the entire image at high resolution, `CPR` formulates perceptual refinement as an instance-adaptive decision process over spatial scale and location. This design enables explicit and controllable compute allocation, allowing the model to progressively focus on diagnostically informative regions without increasing input resolution.

Shared Encoder. Given an input image $X \in \mathbb{R}^{H\times W\times C}$, `CPR` first constructs a global observation by resizing the image to a canonical resolution. The resized image is processed by a shared backbone f_θ, which is reused across all refinement steps. Unlike conventional multi-scale strategies that increase input resolution, `CPR` keeps the input size fixed and instead adapts the perceptual field over the original image, decoupling perceptual granularity from computational cost. This design enables refinement to be achieved through spatial selection rather than resolution scaling, avoiding the quadratic memory growth associated with high-resolution processing. The initial perceptual representation is defined as

$$\mathbf{z}_0 = f_\theta(X_{\text{global}}). \tag{1}$$

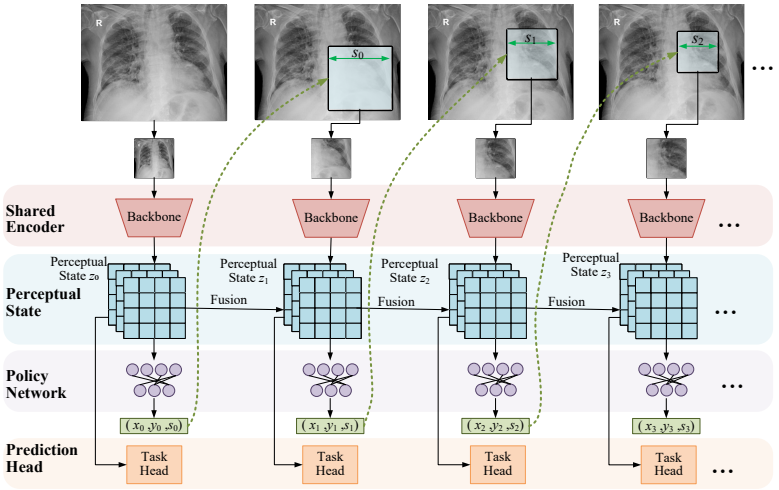

Fig. 2. Overview of the CPR framework. Inspired by human coarse-to-fine visual perception, CPR models medical image interpretation as a sequential refinement process that progressively focuses on localized regions while maintaining constant input resolution, enabling efficient and accurate analysis.

Perceptual State. CPR maintains a latent *perceptual state* that aggregates information across refinement steps, serving as an evolving representation of accumulated evidence. At step t, the localized observation X_t is encoded as

$$\mathbf{Z}_t = f_\theta(X_t). \tag{2}$$

The perceptual state is updated via residual fusion: $\mathbf{H}_{t+1} = \mathbf{H}_t + \mathbf{W}_t\mathbf{Z}_t$ where $\mathbf{H}_t$ denotes the accumulated state and $\mathbf{W}_t$ is a learned projection matrix. This update progressively integrates the global context with localized evidence without increasing input resolution, forming a sequential refinement trajectory. Importantly, the state transition $\mathbf{H}_t \to \mathbf{H}_{t+1}$ encodes the dependency between successive perceptual decisions, allowing future refinement steps to be conditioned on accumulated evidence.

Policy Network. The refinement process is governed by a policy network g_ϕ, which predicts the parameters of the perceptual field conditioned on the current state:

$$\mathbf{a}_t = (x_t, y_t, s_t) \sim g_\phi(\mathbf{H}_t), \tag{3}$$

where (x_t, y_t) denotes the normalized top-left coordinate and s_t denotes the normalized side length of the cropped region. The corresponding crop is extracted from the original image with top-left position (x_tW, y_tH) and spatial size (s_tW, s_tH), and is then resized to the canonical input size before being processed by the shared encoder. These parameters determine *where* to attend and *at what scale* to allocate computation. Throughout the refinement stages, the perceptual scales are encouraged to contract, forming a structured coarse-to-fine trajectory. By conditioning on $\mathbf{H}_t$, the policy learns an instance-adaptive refinement strategy rather than a fixed spatial schedule.

Prediction Head. At each refinement stage, a task-specific prediction head operates on the updated perceptual state:

$$\mathbf{o}_t = h(\mathbf{H}_t), \tag{4}$$

where $h(\cdot)$ denotes the task head.

CPR produces a global prediction $\mathbf{o}^{\text{glob}}$ and refined local predictions $\{\mathbf{o}_t^{\text{loc}}\}_{t=1}^{T}$, allowing progressive error correction across refinement steps and enabling intermediate supervision to stabilize training.

3 Training Objective

Supervised prediction learning. The perceptual backbone (global and local prediction heads) is trained with standard task supervision. Classification loss is applied to the global prediction and all localized predictions:

$$\mathcal{L}_{\text{pred}} = \ell(\mathbf{o}^{\text{glob}}, y) + \sum_{t=1}^{T} \ell(\mathbf{o}_t^{\text{loc}}, y). \tag{5}$$

Prediction consistency. To stabilize progressive refinement, the final localized prediction $\mathbf{o}_T^{\text{loc}}$ is treated as a soft reference. Intermediate predictions are aligned with it via KL divergence:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{pred}} + \alpha \mathcal{L}_{\text{cons}}, \quad \mathcal{L}_{\text{cons}} = \text{KL}(\mathbf{o}^{\text{glob}} \,\|\, \mathbf{o}_T^{\text{loc}}) + \sum_{t=1}^{T-1} \text{KL}(\mathbf{o}_t^{\text{loc}} \,\|\, \mathbf{o}_T^{\text{loc}}). \tag{6}$$

Refinement regularization. To encourage structured coarse-to-fine behavior, we introduce two reward regularizers:

$$\phi_t^{\text{scale}} = \max(0, s_t - s_{t-1}), \quad \phi_t^{\text{rep}} = \mathbb{I}[\hat{\mathbf{a}}_t = \hat{\mathbf{a}}_{t-1}], \tag{7}$$

where s_t denotes the perceptual scale and $\hat{\mathbf{a}}_t$ is the discretized refinement action. The scale regularizer penalizes non-decreasing perceptual scales, while the repetition regularizer discourages consecutive identical refinement actions. The final shaped reward is defined as $\tilde{r}_t = r_t - \lambda_{\text{scale}}\phi_t^{\text{scale}} - \lambda_{\text{rep}}\phi_t^{\text{rep}}$.

Progressive refinement as reinforcement learning. The refinement parameters are generated by a policy network $g_\phi(\mathbf{a}_t \mid \mathbf{H}_t)$, where $\mathbf{H}_t$ denotes the accumulated perceptual state and $\mathbf{a}_t$ encodes the region extraction parameters at step t. The policy is optimized using Proximal Policy Optimization (PPO), treating perceptual refinement as a sequential decision-making problem over spatial scale and location. We define a step-wise refinement reward based on the reduction in task loss: $r_t = \ell(\mathbf{o}_t, y) - \ell(\mathbf{o}_{t+1}, y)$ where $\mathbf{o}_0 = \mathbf{o}^{\text{glob}}$ and $\mathbf{o}_t = \mathbf{o}_t^{\text{loc}}$ for $t \geq 1$.

4 Experiments

4.1 Experiment Setup

Datasets and Baselines. We evaluate CPR on five publicly available medical imaging benchmarks : (1) **Shenzhen Chest X-ray** [9], (2) **COVID-19 Radiography Database** [3], (3) **BUSI Breast Ultrasound** [1], (4) **BreaKHis**

400× [23], and (5) **EyePACS-AIROGS** [11]. These datasets span X-ray, ultrasound, histopathology, and retinal fundus photography, with substantial variation in native resolution. We compare CPR with convolutional and transformer backbones (ResNet-50, ViT-S, Swin-S) under identical training settings. We further evaluate against recent medical imaging methods, including feature fusion (Adaptive NN, OverLock, Hiera, GLNet), feature alignment (mammo-CLIP, MGCA), and multi-instance learning (GABMIL, PAMIL).

Implementation Details. Models are trained for 100 epochs using AdamW with batch size 64. CPR and multi-instance learning methods operate on dataset-adaptive high resolutions (up to 2048^2, 1024^2, and 512^2), resized according to native image scales, while other baseline models use fixed input size.

Table 1. Comparison between CPR and vanilla backbones using ResNet-50, ViT-Small and Swin-Small. NA: no valid result due to OOM; OOM: out-of-memory.

Dataset	ResNet-50			ViT-S			Swin-S		
	(224)	(384)	CPR	(224)	(384)	CPR	(224)	(448)	CPR
Shenzhen	87.88	89.39	**90.15**	76.52	86.36	**90.15**	88.64	NA	**90.91**
COVID-19	97.10	97.28	**98.25**	93.30	98.07	**98.67**	96.20	NA	**98.43**
BUSI	91.05	91.10	**92.01**	89.14	93.33	**94.89**	90.73	NA	**92.97**
BreaKHis	94.31	96.70	**98.35**	97.06	98.53	**99.27**	96.88	NA	**98.53**
EyePACS	94.16	94.68	**94.94**	94.03	94.42	**95.19**	94.16	NA	**94.81**
Memory (G)	3.1	12.1	5.7	3.5	15.8	8.1	8.3	OOM	12.7

4.2 Main Results

Effectiveness of CPR. Table 1 compares CPR with ResNet-50, ViT-S, and Swin-S. CPR consistently outperforms fixed-resolution baselines, showing that gains stem from coarse-to-fine refinement rather than backbone choice. Naive resolution scaling (ResNet-50/ViT-S at 384^2) provides limited benefit. Due to its window-based architecture requiring strict resolution divisibility, Swin-S cannot adopt 384^2 and instead uses 448^2, which results in OOM. In contrast, CPR achieves higher accuracy with lower memory consumption; for example, it improves accuracy by 3.79% while reducing peak GPU memory by 48.73% compared to ViT-S on Shenzhen.

State-of-the-Art Comparisons. Table 2 provides a comprehensive evaluation of CPR against competitive baselines across five public datasets. CPR attains the highest classification accuracy on all benchmarks, consistently surpassing established methods such as Adaptive NN and OverLoCK. Specifically, CPR outperforms the second-best method, Adaptive NN, by margins of 2.27%, 0.24%, 1.92%, 0.45%, and 1.42% on the Shenzhen, Covid, BUSI, BreaKHis, and EyePACs datasets, respectively. Furthermore, CPR maintains a minimal memory footprint of 8.1 G, matching the most efficient baselines while processing images at their native, dataset-adaptive resolutions.

Table 2. Classification accuracy (%) and memory consumption across five public datasets. The best results are in **bold**, and the second-best are underlined.

	Venue	Shenzhen	Covid	BUSI	BreaKHis	EyePACS	Mem. (G)
Mean Resolution	–	2699×2794	1302×977	615×501	700×459	512×512	–
Adaptive NN [26]	NMI'2025	87.88	98.43	92.97	98.72	93.77	**8.1**
OverLoCK [13]	CVPR'2025	89.39	97.89	92.97	97.98	94.03	23.8
Hiera [19]	ICML'2023	85.61	98.19	91.37	93.76	94.81	**8.1**
GLNet [29]	NIPS'2024	89.39	98.43	92.33	98.53	94.16	12.4
mammo-CLIP [6]	MICCAI'2024	88.64	97.52	85.62	86.97	94.29	10.5
MGCA [14]	EAAI'2025	89.39	97.89	92.01	97.43	94.42	8.3
GABMIL [10]	MICCAI'2025	86.36	96.98	84.66	91.19	94.42	8.3
PAMIL [28]	CVPR'2024	84.85	94.63	83.07	90.64	95.06	11.3
CPR	–	**90.15**	**98.67**	**94.89**	**99.27**	**95.19**	**8.1**

Note: NMI denotes *Nature Machine Intelligence*; EAAI denotes *Engineering Applications of Artificial Intelligence*.

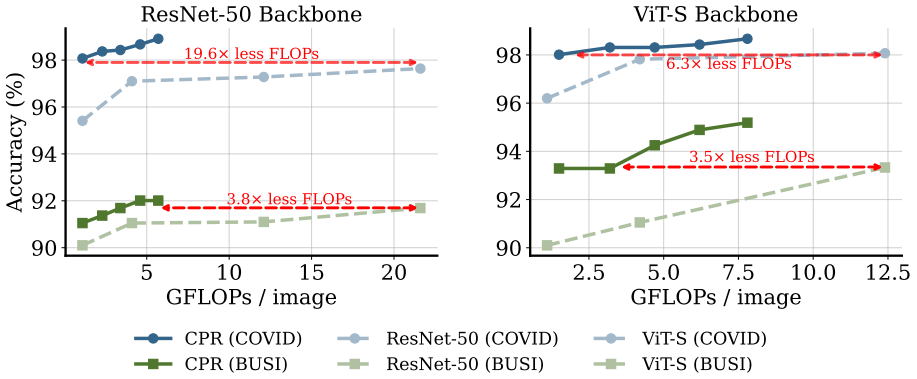

Fig. 3. Accuracy–efficiency trade-offs of CPR (solid) versus fixed-resolution baselines (dashed) on COVID-19 and BUSI. Baselines use resolutions $\{112, 224, 384, 512\}$ for ResNet-50 and $\{112, 224, 384\}$ for ViT-S (higher settings may exceed memory). Red dashed arrows denote GFLOPs reduction at comparable accuracy.

Efficiency. As illustrated in Figure 3, CPR significantly shifts the accuracy and efficiency trade-off by decoupling high-resolution performance from computational cost. By selectively refining salient regions rather than processing full images uniformly, CPR achieves a 19.6× reduction in GFLOPs on the COVID-19 dataset with a ResNet-50 backbone while maintaining comparable accuracy to high-resolution baselines. This efficiency extends across architectures; for the ViT-S backbone, CPR yields a 6.3× and 3.5× FLOP reduction on COVID-19 and BUSI, respectively. Furthermore, as shown in Figure 1, CPR maintains a constant GPU memory footprint regardless of input resolution, effectively bypassing the

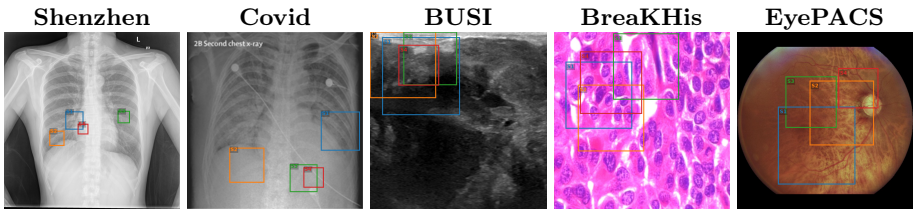

Fig. 4. Progressive refinement boxes (Step1–Step4) on samples per dataset.

quadratic memory growth and Out-of-Memory (OOM) constraints that limit standard models like ViT-S at resolutions beyond 384×384.

Table 3. Ablation study on scale control designs. Base denotes the fine-tuning from pre-trained ViT-S; Base+(x_t, y_t) uses fixed scale (s_t=1); Base+(x_t, y_t, s_t) predicts scale adaptively; Base+$(x_t, y_t, s_t) + \phi_t^{\text{scale}}$ adds explicit scale control; `CPR` denotes Base+$(x_t, y_t, s_t) + \phi_t^{\text{scale}} + \phi_t^{\text{rep}}$.

Dataset	Base	Base+ (x_t, y_t)	Base+ (x_t, y_t, s_t)	Base+ $(x_t, y_t, s_t) + \phi_t^{scale}$	`CPR`
Shenzhen	86.36	87.12	87.12	89.39	**90.15**
COVID-19	96.35	96.98	97.52	98.19	**98.67**
BUSI	92.70	93.33	93.33	93.65	**94.89**
BreaKHis	97.25	97.98	98.35	98.53	**99.27**
EyePACS	91.95	92.47	93.25	94.94	**95.19**

4.3 Additional Analysis

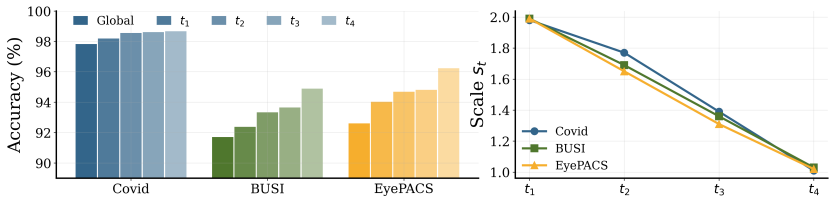

Fig. 5. Qualitative visualization of progressive refinement. (Left:) Accuracy from global to progressive local refinement. (Right:) Scale trend under refinement steps.

Visualization. Fig. 4 shows that `CPR` progressively refines perceptual fields across datasets, moving from global context to discriminative regions. This coarse-to-fine process reduces redundant computation and improves spatial efficiency. The refinement boxes consistently contract toward clinically relevant structures, indicating stable and meaningful scale adaptation.

Ablations. Table 3 compares different scale control formulations. Using only (x_t, y_t) already improves over the backbone but remains inferior, indicating that spatial localization alone is insufficient. Predicting (x_t, y_t, s_t) yields modest gains on some datasets yet is unstable (e.g., Shenzhen shows no improvement over (x_t, y_t)). Introducing the explicit scale-control term $(x_t, y_t, s_t) + \phi_t^{\text{scale}}$ consistently boosts performance. The full `CPR` model achieves the best results across all datasets, validating structured scale regulation and coordinated refinement.

Figure 5 further illustrates progressive refinement in terms of accuracy and scale. As shown in (a), accuracy steadily increases from the global stage to later refinement steps on COVID-19, BUSI, and EyePACS, confirming that each

stage contributes complementary discriminative evidence. In (b), the predicted scale monotonically decreases across steps, reflecting the intended coarse-to-fine mechanism. Together, these findings verify that CPR concentrates computation on informative regions while consistently improving predictive accuracy.

Table 4. Additional control variants on BUSI using ViT-S.

Method	Acc. (%)
Base	92.70
Random Crop	93.29
Fixed Coarse-to-Fine	93.61
Deterministic Attention	94.25
No-Chain CPR	94.25
CPR	**94.89**

Table 5. Comparison with related adaptive architectures. (%)

Method	BUSI	COVID-19
RA-CNN [5]	88.82	98.25
SSPS/M2Former [15]	93.61	95.71
Focus Longer to See Better [22]	90.10	95.59
TNet [18]	94.25	98.37
CPR	**94.89**	**98.67**

Additional controls, baselines, and stability. Table 4 shows that CPR outperforms random cropping, fixed coarse-to-fine scheduling, deterministic attention, and No-Chain CPR, confirming that the improvement benefits from both adaptive policy learning and chained state accumulation. Table 5 further shows that CPR surpasses related iterative/adaptive architectures on BUSI and COVID-19. Separately, to assess policy-learning stability, we repeat CPR with three random seeds and obtain $98.64 \pm 0.17\%$ on COVID-19 and $95.16 \pm 0.30\%$ on BUSI, indicating stable training. Failure cases mainly involve diffuse or weakly contrasted lesions, where over-contracted crops may discard useful context, or salient non-lesion structures that may mislead the policy.

5 Conclusion

We present Chained Perceptual Refinement (CPR), a resolution-aware framework that reformulates high-resolution medical image analysis as a structured, coarse-to-fine reasoning process. By dynamically attending to informative regions via an adaptive policy, CPR sidesteps the prohibitive memory growth typically associated with high-resolution architectures. Our evaluation across five diverse public benchmarks demonstrates that CPR consistently surpasses both fixed-resolution baselines and state-of-the-art multi-scale methods, all while maintaining a constant, hardware-efficient memory footprint.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Grant No. 62501295), the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. XK0020923211), and the Jiangsu Provincial Distinguished Postdoctoral Program (2024ZB490).

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

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (Feb 2020)
2. Ashman, K., Zhuge, H., Shanley, E., Fox, S., Halat, S., Sholl, A., Summa, B., Brown, J.Q.: Whole slide image data utilization informed by digital diagnosis patterns. Journal of Pathology Informatics **13**, 100113 (2022)
3. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al-Emadi, N., Reaz, M.B.I.: Can AI help in screening Viral and COVID-19 pneumonia? IEEE Access **8**, 132665–132676 (2020)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al.: An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021)
5. Fu, J., Zheng, H., Mei, T.: Look Closer to See Better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on CVPR. pp. 4438–4446 (2017)
6. Ghosh, S., Poynton, C.B., Visweswaran, S., et al.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In: MICCAI 2024. pp. 632–642. Springer Nature Switzerland (2024)
7. Haque, M.I.U., Dubey, A.K., Danciu, I., Justice, A.C., Ovchinnikova, O.S., Hinkle, J.D.: Effect of image resolution on automated classification of chest x-rays. Journal of Medical Imaging **10**(4), 044503–044503 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2016)
9. Jaeger S, K.A.: Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging **33**(2), 233–245 (2014)
10. Keshvarikhojasteh, H., Tifrea, M., et al.: A spatially-aware multiple instance learning framework for digital pathology. arXiv preprint arXiv:2504.17379 (2025)
11. Kiefer, R., Abid, M., Ardali, M.R., et al.: Automated fundus image standardization using a dynamic global foreground threshold algorithm. In: International Conference on Image, Vision and Computing. pp. 460–465. IEEE (2023)
12. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
13. Lou, M., Yu, Y.: OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. arXiv preprint arXiv:2502.20087 pp. 1–15 (2025)
14. Lu, S.Y., Zhu, Z., Zhang, Y.D., et al.: Tuberculosis and pneumonia diagnosis in chest x-rays by large adaptive filter and aligning normalized network with report-guided multi-level alignment. Engineering Applications of Artificial Intelligence **158**, 111575 (2025)
15. Moon, J., Lee, J., Lee, Y., et al.: M2Former: Multi-scale patch selection for fine-grained visual recognition. arXiv preprint arXiv:2308.02161 (2023)
16. Murugesan, S.S., Velu, S., et al.: Neural Networks Based Smart E-Health Application for the Prediction of Tuberculosis Using Serverless Computing. IEEE Journal of Biomedical and Health Informatics **28**(9), 5043–5054 (2024)
17. Pang, Y., Zhao, X., Xiang, T.Z., et al.: Zoomnext: A unified collaborative pyramid network for camouflaged object detection. IEEE TPAMI (2024)
18. Papadopoulos, A., Korus, P., Memon, N.: Hard-attention for scalable image classification. In: Advances in NIPS. vol. 34, pp. 14694–14707 (2021)

19. Ryali, C., Hu, Y.T., Bolya, D., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. ICML (2023)
20. Sabottke, C.F., Spieler, B.M.: The effect of image resolution on deep learning in radiography. Radiology: Artificial Intelligence **2**(1), e190015 (2020)
21. Serrão, M.K.M., Costa, M.G.F., Fujimoto, L.B.M., Ogusku, M.M., Costa Filho, C.F.F.: Automatic bright-field smear microscopy for diagnosis of pulmonary tuberculosis. Comput Biol Med **172** (2024)
22. Shroff, P., Chen, T., Wei, Y., et al.: Focus Longer to See Better: Recursively refined attention for fine-grained image classification. In: CVPR. pp. 960–961 (2020)
23. Spanhol, F.A., Oliveira, L.S., Petitjean, C., et al.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016)
24. Vats, S., Sharma, V., Singh, K., et al.: Incremental learning-based cascaded model for detection and localization of tuberculosis from chest x-ray images. Expert Systems with Applications **238** (2024)
25. Wang, C., Piao, S., Huang, Z., Gao, Q., Zhang, J., Li, Y., Shan, H.: Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information. Med Image Anal **91** (2024)
26. Wang, Y., Yue, Y., et al.: Emulating human-like adaptive vision for efficient and flexible machine visual perception. Nature Machine Intelligence pp. 1–19 (2025)
27. Yu, L., Liu, J., Wu, Q., Wang, J., Qu, A.: A Siamese-Transport Domain Adaptation Framework for 3D MRI Classification of Gliomas and Alzheimer's Diseases. IEEE Journal of Biomedical and Health Informatics **28**(1), 391–402 (2024)
28. Zheng, T., Jiang, K., Yao, H.: Dynamic policy-driven adaptive multi-instance learning for whole slide image classification. In: CVPR (2024)
29. Zhu, L., Wang, X., Zhang, W., et al.: Revisiting the integration of convolution and attention for vision backbone. Advances in NIPS (2024)