



OPEN Adaptive patch sampling and location-aware reasoning for whole body PET-CT multi-organ segmentation

Junha Park¹, Arthur Cho² & Hae-Jeong Park^{1,2,3,4,5,6}✉

Patch-wise learning is a common strategy for training neural networks on large-scale dense prediction problems, yet existing approaches assume uniform or fixed sampling distributions. This assumption is suboptimal when learning difficulty varies spatially and evolves with the model state during optimization. We reformulate patch-wise learning as a dynamic computation allocation problem and propose an adaptive patch sampling (APS) algorithm that learns where to sample by constructing model-state-dependent sampling distributions from voxel-wise uncertainty and prediction error. To learn what contextual information is encoded within sampled patches, we introduce a patch encoding (PE) block that infers implicit location information and modulates feature representations through context-dependent channel-wise attention, without relying on explicit spatial coordinates. Experiments on whole-body multi-organ PET-CT segmentation demonstrate faster convergence and consistent performance gains, with external validation on Synapse dataset confirming robustness. Mechanistic analyses of learning dynamics further characterize sampling behavior induced by APS and representation modulation driven by the PE block through attention analysis and causal channel pruning. Overall, this work contributes an efficient learning strategy for patch-wise training and provides insight into how dynamic sampling and contextual conditioning influence optimization in large-scale dense prediction tasks.

Patch-wise learning is a widely adopted strategy for training neural networks on large-scale dense prediction problems, particularly when memory and computational constraints prohibit full-volume optimization. This paradigm has been extensively used in volumetric segmentation tasks, including multi-modal PET-CT analysis for lesion detection and healthy organ segmentation^{1–6}. In patch-wise learning, high-resolution images are decomposed into fixed-size patches to enable efficient training of convolutional and transformer-based models^{7–10}. Despite its practical success, most patch-wise training pipelines rely on uniform or fixed non-uniform sampling schemes that implicitly assume static informativeness of local regions throughout optimization.

From an optimization perspective, this assumption is often violated. In dense prediction tasks, prediction error and model uncertainty vary substantially across spatial regions and evolve dynamically as model parameters are updated. Consequently, uniform sampling leads to inefficient allocation of computation by sampling excess low-informative regions. This issue is particularly pronounced in problems characterized by heterogeneous structures and complex boundaries, where different regions contribute unevenly to the optimization landscape^{11,12}. Reducing redundant sampling is therefore critical not only for accelerating convergence but also for improving computational and energy efficiency in large-scale training.

Several prior studies have attempted to address this issue by introducing non-uniform patch sampling strategies computed before training^{12–14}. These approaches define fixed sampling distributions based on organ size imbalance¹², spatial prior obtained from registration or atlas construction¹³, or explicit prioritization of predefined target regions¹⁴. While effective in specific settings, such strategies rely on static priors derived from the dataset rather than the evolving state of the model. As a result, they fail to adapt to changes in the loss

¹Yonsei University College of Medicine, Seoul, South Korea. ²Department of Nuclear Medicine, Yonsei University College of Medicine, Seoul, South Korea. ³Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, South Korea. ⁴Department of Cognitive Science, Yonsei University, Seoul, South Korea. ⁵Institute of Human Complexity and Systems Science, Yonsei University, Seoul, South Korea. ⁶Epilepsy Research Institute, Brain research institute, Yonsei University, Seoul, South Korea. ✉email: parkhj@yonsei.ac.kr

landscape during optimization. This limitation motivates a reformulation of patch sampling as an online, model-state-dependent process.

In this work, we propose to integrate adaptive sampling directly into patch-wise learning by constructing sampling distributions from voxel-wise prediction error and uncertainty estimated during training. By coupling data selection with the evolving optimization dynamics, adaptive patch sampling reallocates computation toward informative regions without requiring precomputed priors or rejection-based sampling schemes. This design enables smooth, online adaptation of the sampling policy throughout training.

Beyond sampling, patch-wise learning also raises a fundamental question of representation: how contextual information is encoded when global spatial structure is fragmented into local patches. Since patch sampling determines the spatial location of patch centers, patch-level reasoning inherently depends on contextual cues related to location. However, commonly used segmentation architectures, including convolutional neural networks and transformer-based models^{8–10,15,16}, lack explicit mechanisms for reasoning about spatial context when operating on isolated patches. Although explicit coordinate encoding or landmark-based approaches can provide positional information, such methods are often unreliable in practice due to variability in acquisition ranges, anatomical proportions, and the need for additional annotations¹⁷.

To address this challenge, we propose to infer implicit contextual cues directly from the semantic content of patches. Specifically, we hypothesize that the presence or absence of specific structures within a patch provides sufficient information to guide context-aware reasoning. We introduce a patch encoding (PE) block that integrates such implicit contextual information into feature representations via channel-wise attention mechanisms^{18–21}, without relying on explicit spatial coordinates.

The joint focus on *where* patches are sampled and *what* contextual information they encode naturally enables analysis of learning dynamics. By explicitly modeling the sampling process and contextual conditioning, we can examine how computation is allocated during training and how internal representations evolve in response to adaptive sampling. This perspective allows algorithm-level analysis of patch-wise learning behavior beyond post hoc performance evaluation.

This study makes three contributions. First, we propose adaptive patch sampling (APS), which reformulates patch-wise training as adaptive sampling with a dynamic, model-state-dependent distribution. Second, we introduce a patch encoding (PE) block that enables implicit, context-aware representation learning via channel-wise attention. Third, we provide a systematic analysis of sampling dynamics and representation modulation, offering mechanistic insight into how adaptive sampling and contextual conditioning influence optimization in patch-based learning.

We demonstrate the effectiveness of the proposed framework on whole-body PET-CT segmentation as a challenging testbed for large-scale patch-wise learning, showing improved segmentation performance, optimization efficiency, and enhanced interpretability across multiple structures.

Algorithm

Algorithm overview

We present a learning framework that explicitly addresses two fundamental questions in patch-wise training: *where* to allocate computation during optimization and *what* contextual information should be encoded within sampled patches. The proposed framework consists of two tightly coupled components, adaptive patch sampling (APS) and a patch encoding (PE) block, integrated into a standard patch-based training pipeline.

Given a high-resolution input volume, training proceeds by iteratively sampling fixed-size patches and updating model parameters using stochastic gradient descent. Unlike conventional approaches that sample patch locations uniformly or from fixed distributions, APS dynamically determines the sampling distribution at each training stage. Specifically, APS estimates voxel-wise prediction error and uncertainty from the current model state and constructs a model-state-dependent importance distribution over spatial locations. Patch centers are then sampled from this distribution, enabling the training process to preferentially focus on regions that contribute most to the optimization objective. This adaptive sampling policy is updated online throughout training and does not require precomputed prior, rejection-based sampling, or additional preprocessing.

Once patches are sampled, they are processed by a segmentation network augmented with the proposed PE block. The PE block addresses the loss of global context inherent to patch-wise learning by inferring implicit contextual cues from the semantic content of each patch. It predicts organ existence patterns within the patch and transforms this information into a channel-wise attention vector, which modulates intermediate feature representations. Through this mechanism, the network conditions its predictions on inferred contextual information without relying on explicit spatial coordinates or external annotations.

The interaction between APS and the PE block forms a closed-loop learning system. APS influences *where* the model observes data by reallocating sampling effort based on the evolving optimization landscape, while the PE block influences *how* sampled information is represented by conditioning features on inferred context. Together, these components improve sample efficiency, accelerate convergence, and enable analysis of learning dynamics by exposing both sampling behavior and representation modulation during training.

An overview of the framework is illustrated in Fig. 1, highlighting the adaptive sampling process, the integration of the PE block within the network, and the feedback loop between model state and data selection.

Adaptive patch sampling as a biased random walk on image patches

Let $L(\theta)$ denote the training loss parameterized by θ . Under gradient descent, parameters are updated as $\theta_{t+1} = \theta_t - \lambda \nabla L_{\theta_t}$ where t indexes the training step and λ denotes the learning rate. In patch-wise training, gradients are estimated from a mini-batch of image patches $\{x_i\}_{i=1}^B$ sampled from the image domain Ω , yielding a monte carlo estimator $\nabla L_{\theta} \approx \frac{1}{|\Omega| \cdot B} \times \sum_{i=1}^B w(x_i) \nabla L(x_i)$ where $p(x_i)$ and $w(x_i)$

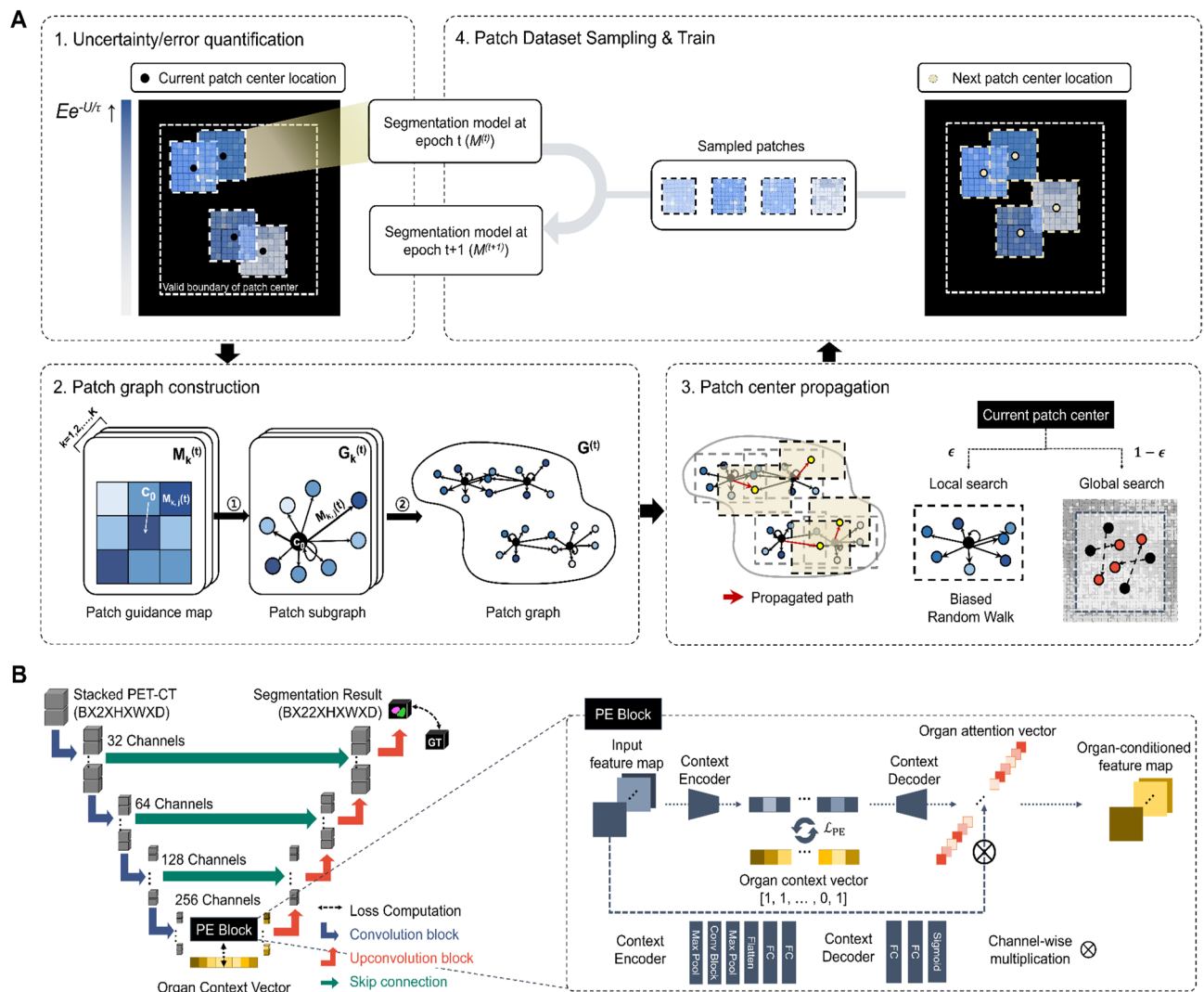


Fig. 1. Overview of the proposed patch-wise learning framework. **(A)** Adaptive patch sampling (APS) formulated as a dynamic importance sampling, where patch selection is guided by model-state-dependent uncertainty and prediction error. **(B)** Patch encoding (PE) block that conditions feature representations on inferred contextual cues through channel-wise attention, enabling location-aware reasoning without explicit spatial coordinates.

define the sampling strategy. Adaptive sampling selects non-uniform $p(x_i)$ but uniformly weight $w(x_i)$ thus naturally modifying the domain Ω towards those with high sampling probabilities²².

Existing non-uniform patch sampling methods typically rely on fixed distributions computed before training, based on organ size, priors from spatial registration, or target regions^{12–14}. However, segmentation difficulty evolves during training as model parameters change. We therefore model patch sampling as a dynamic process conditioned on the model state, formalized as a biased random walk over image patches²³.

A 3D patch centered at coordinate (x, y, z) is defined as $P(x, y, z) = \{I(x', y', z') | x - \lfloor \frac{p_x-1}{2} \rfloor \leq x' \leq x + \lfloor \frac{p_x-1}{2} \rfloor, y - \lfloor \frac{p_y-1}{2} \rfloor \leq y' \leq y + \lfloor \frac{p_y-1}{2} \rfloor, z - \lfloor \frac{p_z-1}{2} \rfloor \leq z' \leq z + \lfloor \frac{p_z-1}{2} \rfloor\}$, where (p_x, p_y, p_z) denotes the 3D patch size and I is the original image.

At each epoch t , we construct a patch-set $S^{(t)} = \{P_k^{(t)} | k \in \{1, 2, \dots, K\}\}$ for each training image by sampling K patches. Each patch in a patch-set induces a patch-subgraph whose nodes correspond to voxels, with directed edges from the patch center to all voxels in the patch weighted by a voxel-wise guidance map $M_k^{(t)}$.

The union of all patch subgraphs forms a patch-graph $G^{(t)}$ to guide patch sampling process by performing biased random walk on this graph: $Pr(C^{(t+1)}_k = x | C^{(t)}_k = v) = \frac{M_{v,x}}{Z}$, where $Z = \sum_{x \in N(v)} M_{v,x}$ is a normalizing constant computed for the neighboring voxels $N(v)$ reachable from the current patch center (Fig. 1A).

Approximation of optimal sampling distribution with voxel-wise model predictions

Importance sampling theory in deep neural networks shows that the optimal sampling distribution is proportional to the per-sample gradient norm^{24–26}. However, directly computing the full gradient norm is impractical in patch-wise training because only a limited number of patches $P^{(t)} \in S^{(t)}$ are observed at a time t , whereas an ideal estimate would require evaluating the gradient norm of all candidate patch centers across the entire image I . To address this, we decompose the optimal patch sampling distribution as a weighted sum of two components: a global component of uniform distribution and a local component of adaptive sampling distribution, controlled by a weight parameter ϵ (Supplementary Fig. 1).

Let us consider a set of all valid voxels V_{valid} to define a global component, where a voxel is valid if its minimum margin from the image boundary exceeds half of the patch size. To define a local component, let $M_k^{(t)}$ be the local guidance map over candidate voxels within the current patch $P_k^{(t)}$ at epoch t , then the patch sampling probability is defined as:

$$Pr\left(P^{(t+1)} = P_{xyz}|M_k^{(t)}\right) = \epsilon \cdot p_{local} + (1 - \epsilon) \cdot p_{global} = \epsilon \cdot \frac{M_{k,xyz}^{(t)}}{\sum_{(x,y,z) \in P^{(t)}} M_{k,xyz}^{(t)}} + (1 - \epsilon) \cdot \frac{1}{|V_{valid}|} \quad (1)$$

Ideally, the optimal guidance map $M_{k,xyz}^{(t)}$ would be proportional to the gradient norm obtained by back-propagating the loss over entire patch P_{c_0} with the corresponding label. However, explicitly computing $M_{k,xyz}^{(t)}$ is computationally infeasible as it would require exhaustive forward and backward passes over patches centered at every possible voxel in the image. To avoid this prohibitive cost, we leverage the information already available during training. Instead of recomputing gradients for all potential patch centers, we reuse the uncertainty map $U_k^{(t)}$ and prediction error map $R_k^{(t)}$ produced for each sampled patch $P_k^{(t)}$ at epoch t . Recent theoretical analysis under voxel-wise cross-entropy loss shows that the full parameter gradient norm is upper-bounded by the gradient norm at the output layer, which admits a closed-form expression depending only on the dissimilarity between the predicted probability distribution and the true label²². This result motivates the use of voxel-wise uncertainty map $U_k^{(t)}$ and error map $R_k^{(t)}$ as efficient surrogates for the true gradient norm. Both uncertainty and error maps can be computed directly from a single forward pass of the segmentation model $f^{(t)}(P, c)$ from the patch P and class c at epoch t . Formally, each map is defined as:

$$\begin{aligned} \Delta_{k,xyz}^{(t)} &= \max_{c \in C} f^{(t)}(P_k^{(t)}, c)_{xyz} - \max_{c \in C \setminus c^+} f^{(t)}(P_k^{(t)}, c)_{xyz}, \quad U_{k,xyz}^{(t)} = \exp\left(-\Delta_{k,xyz}^{(t)}/\tau\right), \\ R_{k,xyz}^{(t)} &= \{0, \text{ if } \operatorname{argmax}_c f^{(t)}(P_k^{(t)}, c)_{xyz} = c_{ground\ truth}, 1, \text{ otherwise}\}, \quad M_k^{(t)} = R_k^{(t)} \odot U_k^{(t)} \end{aligned} \quad (2)$$

Here, c denotes a class in the set of available organ classes C , and c^+ is the class with the highest voxel confidence at coordinate (x, y, z) . The notation $C \setminus c^+$ refers to all classes in the set C , excluding c^+ . The model uncertainty map $U_k^{(t)}$ is computed by applying negative exponentiation of the difference $\Delta_k^{(t)}$ between the top two SoftMax probabilities divided by the temperature $\tau = 100$. The error map $R_k^{(t)}$ assigns 1 to incorrectly predicted voxels and 0 to correctly predicted ones within the patch $P_k^{(t)}$. We examined the distribution of uncertainty-error map values at the first epoch of training when the network is initialized and chose $\tau = 100$ to reduce stochastic variation in the early training dynamics (Supplementary Fig. 2).

In practice, the implemented value of ϵ is slightly downward biased compared to the intended value, due to the case when an adaptively chosen patch center lies outside the valid boundary and is replaced by a uniform sampling (corresponding possibility is negligible in most of the coordinates throughout the whole image). When $\epsilon = 0$, the algorithm defaults to global uniform patch sampling, rejecting all locally proposed samples. When $\epsilon = 1$, all samples are drawn via a biased random walk on the patch graph, except when the proposed patch location lies outside the valid patch center boundary.

Patch encoding via implicit location-aware channel attention

To incorporate implicit locational context, we introduce a patch encoding (PE) block at the bottleneck of the baseline U-Net, hereafter referred to as U-Net+PE (Fig. 1B). Given PET and CT images I_{PET} and I_{CT} , the segmentation encoder network f_{Senc} produces a feature map $\hat{y}_i = f_{Senc}(I_{PET}, I_{CT})$ which is subsequently passed to the PE block. The PE block consists of two components, context encoder and context decoder that incorporate location information and produce an enhanced feature map $\hat{y}_i'$.

For the input feature map y_i of size (A, X, Y, Z) , corresponding to (channel, width, height, depth), context encoder maps y_i to an organ context vector $\hat{y}_c = f_{Cenc}(\hat{y}_i) \in R^C$ where $C = 22$ denotes the number of organ classes (including background). Organ context vector $\hat{y}_c$ is aligned with the ground truth binary organ existence label y_c by minimizing the negative log-likelihood loss. Unlike global contextual factors (e.g., arm position, sexual difference), the existence of a specific organ within a patch provides essential contextual information implicitly, without requiring additional annotation. This is feasible because organ existence within a patch can be directly inferred from the segmentation label.

Subsequently, context decoder maps the estimated organ context vector $\hat{y}_c$ to organ attention vector $\hat{y}_a = f_{Dec}(\hat{y}_c) \in R^A$ where $A = 256$ denotes the number of U-Net bottleneck channels. The input feature map $\hat{y}_i$ is modulated via channel-wise multiplication $\hat{y}_i' = \hat{y}_i \otimes \hat{y}_a$. The segmentation decoder network

$f_{S_{dec}}$ performs segmentation using both $\hat{y}_i$ and $\hat{y}_a$ as inputs: $\hat{y}_{seg} = f_{S_{dec}}(\hat{y}_i, \hat{y}_a)$. The organ-conditioned feature map is forwarded to the decoder layer only when the skip connection is replaced by the PE block.

The loss functions for segmentation and classification are combined into a multi-task loss function $L_{mt} = L_{seg}(\hat{y}_{seg}, y_{seg}) + \lambda L_{cls}(\hat{y}_c, y_c)$. The segmentation loss is defined as $L_{seg} = L_{dsc} + \beta L_{ce}$, where L_{dsc} is the Dice score loss and L_{ce} is the cross-entropy loss. The classification loss L_{cls} within the PE block is computed using binary cross-entropy. Each component of the multi-task loss is detailed in the following equation.

$$L_{dsc} = 1 - (2/L) \cdot \sum_{j=1}^L \frac{\sum_{i=1}^N y_{seg,i}^{(j)} \hat{y}_{seg,i}^{(j)}}{\sum_{i=1}^N (y_{seg,i}^{(j)} + \hat{y}_{seg,i}^{(j)})}, L_{ce} = -(1/NL) \cdot \sum_{i=1}^N \sum_{j=1}^L \hat{y}_{seg,i}^{(j)} \log(y_{seg,i}^{(j)}), \quad (3)$$

$$L_{cls} = -(1/L) \cdot \sum_{j=1}^L \sigma(\hat{y}_c^{(j)}) \log(y_c^{(j)})$$

Here, σ denotes a sigmoid function, $y_{seg,i}^{(j)}$ and $\hat{y}_{seg,i}^{(j)}$ represents the ground truth label and continuous values of probabilistic prediction for the organ j at the voxel i , respectively. For all evaluations in this study, we empirically selected $\lambda = 0.2$ and $\beta = 0.05$, following prior multi-task learning approaches^{27,28}.

Methods

Dataset preparation

We used two datasets for validation, an in-house whole-body PET-CT dataset and an external multi-organ segmentation benchmark.

For the in-house dataset, we collected 159 whole-body PET-CT image pairs from healthy individuals undergoing routine clinical check-ups at Severance Hospital. Each image was annotated into 21 organ regions: right lung, left lung, heart, liver, spleen, pancreas, right kidney, left kidney, aorta, stomach, skin, fat, bone, right shoulder muscle, left shoulder muscle, chest muscle, back muscle, psoas muscle, iliacus muscle, right thigh muscle, and left thigh muscle. Annotations were created in 3D Slicer²⁹ using delineation criteria from prior research and our internal guidelines. LDCT images were initially segmented, and these annotations subsequently guided the labeling of the PET images.

All images were preprocessed before training. Voxel sizes were standardized to $4.1 \times 4.1 \times 3$ mm³, and volumes were resized to dimensions within $200 \times 200 \times (263\text{--}352)$. PET images underwent logarithmic intensity transformation, while LDCT images were windowed to the range $(-300, 300)$ Hounsfield units and linearly rescaled to $(0, 255)$.

For external validation, we used the Synapse dataset from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, a widely used CT-based multi-organ segmentation benchmark³⁰. The Synapse dataset contains abdominal CT scans with expert annotations for multiple organs, including the liver, spleen, kidneys, stomach, and pancreas. Although Synapse does not include PET images, it provides an appropriate testbed to evaluate the generalizability of the proposed training strategy beyond PET-CT imaging.

Unless otherwise specified, all method development, ablation studies, and interpretability analyses were conducted using the in-house PET-CT dataset. Performance evaluation on the Synapse dataset was used exclusively for external validation.

Implementation details

All models were implemented in PyTorch³¹ and MONAI³². To evaluate the pure methodological contributions of the proposed APS and PE block, we adopted a standard 3D U-Net¹⁶ as the baseline architecture rather than more sophisticated frameworks such as nnU-Net⁸. This choice allows us to isolate the effects of the proposed sampling and reasoning strategies from confounding factors introduced by automated preprocessing pipelines, architecture self-configuration, and task-specific training heuristics that are tightly integrated in nnU-Net. The baseline 3D U-Net employed uniform patch sampling, with PET and LDCT images stacked along the channel dimension as input. Each convolution block in the encoder, i.e., a building block of a 3D U-Net encoder, consists of two convolutional layers followed by instance normalization³³ and leaky ReLU³⁴ activation. The up-convolution block in the decoder, a building block of the U-Net decoder, consists of a single transposed convolutional operation followed by the same normalization and activation function scheme. The number of channels in the feature maps obtained from convolution and up-convolution blocks follows the sequence 32, 64, 128, 256, 128, 64, 32 (Fig. 1B).

The U-Net + PE model incorporated a PE block at the deepest layer of skip connection from the third encoder block to the first decoder block. The PE block consists of a context encoder and a context decoder. The context encoder begins by downsizing the bottleneck feature map, employing max pooling with a kernel size of 2. A convolution block and adaptive max-pooling follow this step, applied consecutively to reduce the size of the feature map further. The resulting three-dimensional feature map is then flattened and passed through two sequential linear layers. This process generated a 22-dimensional vector, which corresponds to the organ context modelled by the existence of organs. The context decoder comprises two linear layers that project the organ context vector into a higher-dimensional vector space (Fig. 1B). The resultant organ attention vector is tailored to match the dimensionality of the bottleneck channel number. Subsequently, this organ attention vector is subjected to channel-wise multiplication with the input feature map initially fed into the PE block. Through this operation, we obtain a refined feature map conditionally influenced by the patch's contextual information.

For network training, we used the Adam optimizer³⁵ with an initial learning rate 2×10^{-3} and the learning rate was modulated using the CosineAnnealingWarmRestarts scheduler³⁶, tapering between 2×10^{-3} and

2×10^{-5} with a period of 10 epochs. U-Net, U-Net + PE, and Swin-UNETR¹⁰ were trained for 100 epochs with batch sizes of 4, sampling eight patches (uniformly or adaptively) with a size of $64 \times 64 \times 64$ per PET-CT pair to construct a training dataset for each epoch. Inference used a sliding window approach with an overlap ratio of 0.7. Performance was evaluated via 5-fold cross-validation using Dice score³⁷ as a metric. For external validation, we applied the proposed APS and PE block to the Synapse dataset using identical training hyperparameters, except that the patch size was increased to $128 \times 128 \times 128$ for CT images. Given the relatively small number of samples in the dataset ($n = 30$), performance was evaluated using 3-fold cross-validation.

Quantitative experiments and mechanistic analyses

We compared the segmentation performance of the proposed APS algorithm against the uniform sampling baseline for both U-Net and U-Net + PE, sweeping the weighting parameter ϵ from 0.1 to 0.9 with a step of 0.1. For notation, we denote U-Net and U-Net + PE trained with APS as U-Net _{ϵ} and U-Net _{ϵ} + PE by specifying ϵ .

Beyond the quantitative benefit of segmentation performance, we examined the mechanisms underlying our framework's learning of *where* to sample and *what* anatomical context patches represent (Fig. 2A). To interpret where APS concentrates sampling effort, we introduce cumulative sampling frequency (CSF), defined as the total number of times a voxel is included in the K sampled patches per epoch, summed over T training epochs. Formally, CSF value CSF_{xyz} at coordinate (x, y, z) is defined as $CSF_{xyz} = \sum_{t=1}^T \sum_{k=1}^K J_{k,xyz}^{(t)}$, where $J_{k,xyz}^{(t)}$ is an indicator function whether the voxel positioned at (x, y, z) of image I is included in the patch $P_k^{(t)}$ sampled at epoch t . This provides an explicit measure of the model's spatial attention throughout training.

To interpret *what* anatomical information the PE block learns, we investigate whether it implicitly encodes spatial context by predicting the organ's existence (Fig. 2A). We conducted three experiments to evaluate whether (i) U-Net + PE model accurately predicts organ existence, (ii) organ attention vectors of PE block align with the anatomical location of the sampled patches, and (iii) organ attention vectors selectively emphasize anatomically relevant channels in the feature map. For this analysis, we trained a U-Net + PE model to jointly predict the segmentation map and organ existence and compared its organ classification performance with a baseline classification network (CLS). CLS network is defined as a composite($f_{C_{enc}} \circ f_{S_{enc}}$) of U-Net encoder ($f_{S_{enc}}$) and PE block encoder ($f_{C_{enc}}$), trained solely for organ existence prediction (Fig. 2B). We further collected 256-dimensional organ attention vectors from the PE block by sliding $64 \times 64 \times 64$ image patches with a 32-voxel overlap across the whole body. These vectors were projected into 2D t-SNE embeddings to visualize patch center locations in the embedding space (Fig. 2B). Finally, we binarized the continuous values of the 256-dimensional organ attention vector using a threshold of 0.5. This binarization resulted in a pruned version of the bottleneck feature map by zeroing out the channels with the corresponding organ attention value below the threshold. We then evaluated whether this perturbation degrades the model performance, thereby testing the hypothesis that the organ attention vector selectively emphasizes crucial channels of the feature map.

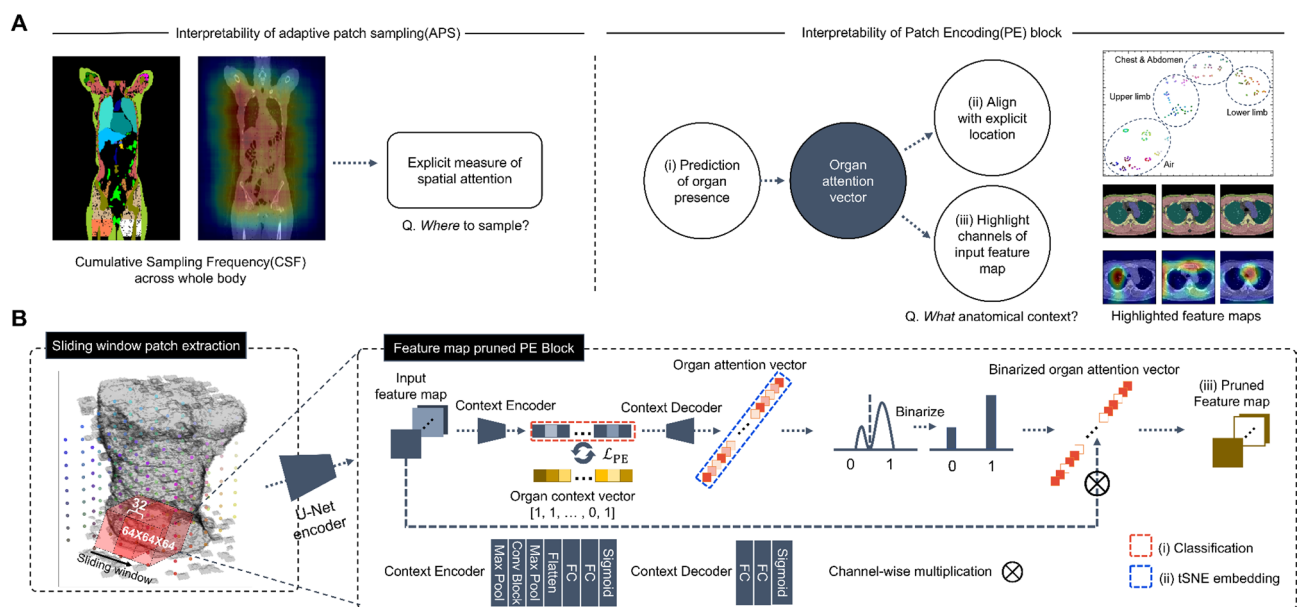


Fig. 2. Interpretability analyses of learning *where* and *what* in patch-wise training. **(A)** Overview of the mechanistic interpretability objectives, analyzing *where* the adaptive patch sampling (APS) allocates training focus and *what* anatomical context is encoded by the patch encoding (PE) block. **(B)** Three complementary analyses used to validate that learned mechanisms: (i) organ existence prediction by the PE block (classification task; red dashed box), (ii) spatial organization of learned organ attention vectors visualized by t-SNE embedding (blue dashed box), and (iii) selective emphasis of information feature map channels assessed through attention-guided channel pruning.

Ethical considerations

All methods were performed in accordance with the relevant guidelines and regulations. All study participants provided informed consent before participating. Ethical approval of this retrospective study was obtained by the Institutional Review Board of Severance Hospital.

Results

Quantitative segmentation performance

Multi-organ segmentation is known to exhibit substantial heterogeneity in difficulty across different organs, and our results reflect this pattern. Across the 21 target organs, baseline performance of the U-Net trained with uniform sampling on PET-CT images ranged from 0.652 to 0.971 in Dice score (Fig. 3A). Specifically, organs such as the liver and lungs achieved high Dice scores, whereas the pancreas showed considerably lower performance.

To assess whether this organ-level heterogeneity reflects intrinsic spatial variation in segmentation difficulty rather than model-specific effects, we compared the organ-level Dice scores obtained from our PET unimodal and PET-CT multimodal approaches with those reported in a previous multi-organ segmentation benchmark³⁸. We observed a strong correlation between methods at the organ level (Spearman $R = 0.975$, $p < 0.001$; Supplementary Fig. 2), suggesting that relative segmentation difficulty across organs is highly consistent across datasets and segmentation approaches. In addition, the PET-CT multimodal approach consistently outperformed the PET unimodal approach, confirming that complementary contrast from PET and LDCT images benefits organ segmentation (Fig. 3A).

Both adaptive patch sampling (APS) and the patch encoding (PE) block improved global segmentation performance averaged across all organs (Fig. 3B), while revealing heterogeneous effects at the organ level (Fig. 3C and D). When APS was applied to the in-house dataset, we observed consistent performance gains over the segmentation model trained with uniform sampling ($\epsilon = 0$). The optimal balancing parameter was $\epsilon = 0.2$ for U-Net and $\epsilon = 0.3$ for U-Net + PE, corresponding to statistically significant improvements of 1.9% and 1.5% in Dice score, respectively, compared with uniform sampling (Wilcoxon signed-rank test, one-sided, $p < 0.05$). We observed consistent performance gains when APS was applied to Swin-UNETR on the in-house dataset (Supplementary Fig. 4), suggesting that APS is generalizable to CNN and Transformer-based segmentation networks.

When APS was applied to the Synapse dataset using the optimal ϵ values determined on the in-house dataset, U-Net _{$\epsilon=0.2$} and U-Net _{$\epsilon=0.3$} + PE outperformed their uniform sampling counterparts by 1.3% and 2.2% Dice score improvements, respectively, demonstrating strong generalizability (Supplementary Fig. 3).

When the PE block was applied to our in-house dataset, U-Net _{ϵ} + PE significantly improved (Wilcoxon signed-rank test, one-sided, $p < 0.05$) the average Dice score compared to U-Net _{ϵ} for all tested ϵ values ranging from 0 to 0.9 (Fig. 3B). U-Net _{$\epsilon=0$} + PE model significantly outperformed baseline U-Net _{$\epsilon=0$} model by 0.7% (Wilcoxon signed-rank test, one-sided, $p < 0.05$) in Synapse dataset, synergizing with APS (U-Net _{$\epsilon=0.3$} + PE) to achieve 3.0% of significant (Wilcoxon signed-rank test, one-sided, $p < 0.05$) overall performance gain compared to the baseline U-Net _{$\epsilon=0$} model (Supplementary Fig. 3).

Convergence behavior and computational efficiency

We next evaluated whether APS accelerates convergence during patch-wise training. Convergence speed was assessed using Page's trend test³⁹, a non-parametric test for detecting monotonic trends in ordered fitness observations. Validation Dice score and validation loss were used as fitness measures. For both U-Net and U-Net + PE trained on the in-house dataset, models trained with optimal APS converged significantly faster than those trained with uniform sampling, as evidenced by validation Dice score ($p < 0.001$) and validation loss ($p < 0.001$; Fig. 3E).

Regarding computational overhead, APS and the PE block increased training time per epoch by only 0.27% and 0.35%, respectively, compared with the baseline U-Net trained with uniform sampling. These negligible overheads indicate that improved convergence and accuracy are achieved without substantial additional computational cost.

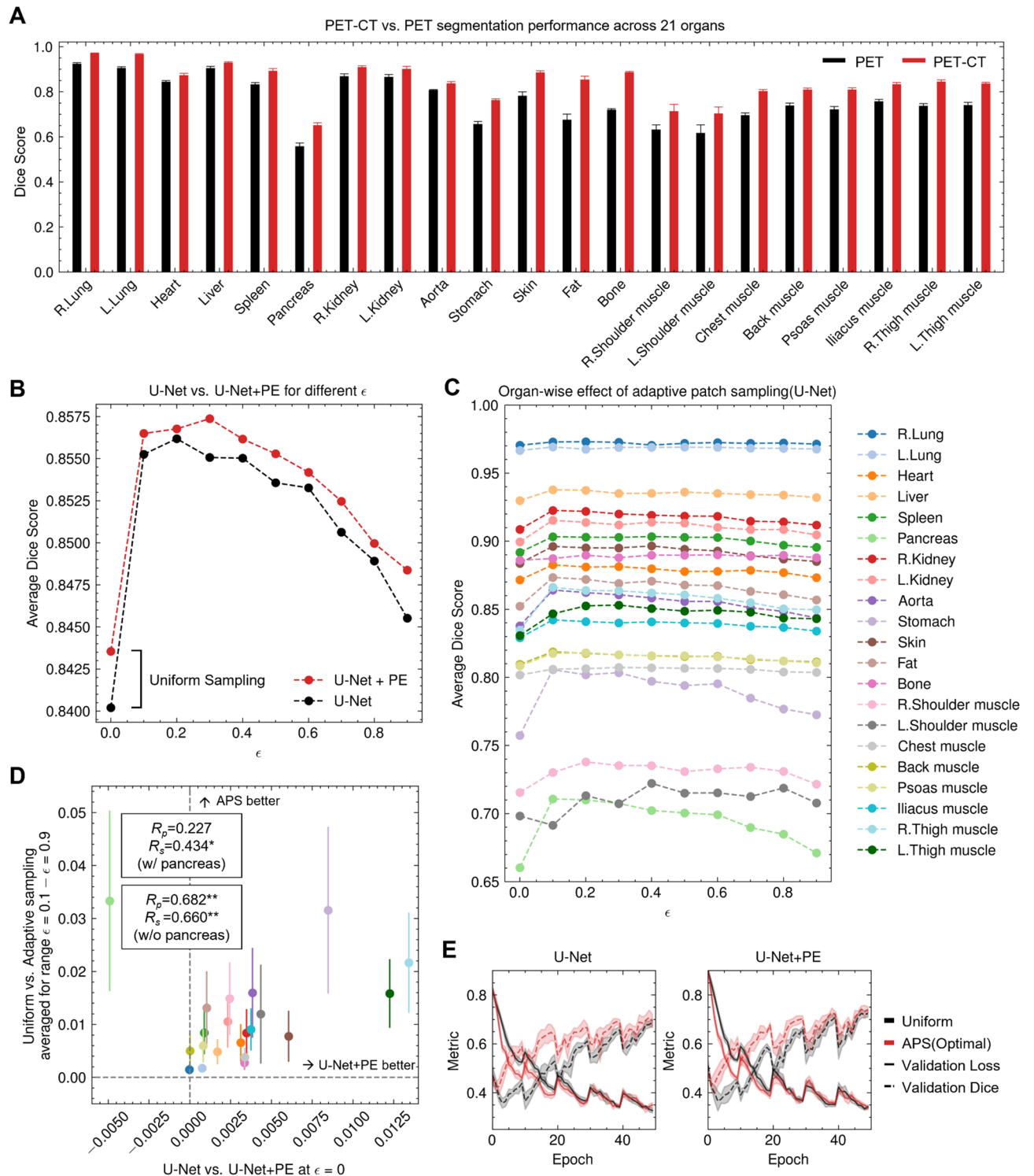
Organ-specific performance analysis

At the organ level, pancreas segmentation exhibited the lowest baseline performance but showed the largest improvement when APS was applied (Fig. 3C and D). We interpret this gain as evidence that pancreas segmentation is under-optimized under uniform sampling due to insufficient exposure during training, which can be effectively addressed through adaptive sampling.

In contrast, segmentation of the stomach and left/right thigh muscles benefited most from the PE block, suggesting that these organs preferentially require contextual information for accurate inference. Joint analysis of performance gains revealed a significant correlation between APS- and PE-induced improvements across organs (Spearman $R = 0.660$, $p < 0.005$, excluding pancreas), while also identifying organ-specific deviations from this trend (Fig. 3D). Organs larger than the patch size, such as the stomach and thigh muscles, are more prone to partial observation, leading to higher variability in appearance and stronger dependence on contextual cues.

Mechanistic interpretability of APS

We validated the mechanisms underlying the proposed framework through a series of interpretability analyses, examining both where APS directs sampling attention and what anatomical context the PE block encodes. We interpret the cumulative sampling frequency (CSF) map as a measure of training-time spatial attention, visualizing regions repeatedly visited during training. For both uniform and adaptive sampling, voxels near the image boundaries exhibited lower CSF values due to reduced inclusion in randomly sampled patches. Beyond



this shared pattern, APS produced markedly different CSF distributions, progressively concentrating sampling on abdominal regions rather than uniformly covering the volume (Fig. 4A).

We further examined the CSF maps under high values of the balancing parameter ϵ values ($\epsilon = 0.7$ and $\epsilon = 0.9$), which were associated with reduced segmentation performance relative to the optimal setting (Fig. 3C). Under these high ϵ values, APS substantially reduced sampling from the lower extremities, suggesting that excessive concentration on a subset of regions may hinder generalization across all organs. This observation highlights the necessity of the balancing parameter ϵ in regulating the trade-off between focused sampling and sufficient spatial coverage.

To quantify organ-specific sampling behavior at the optimal training setting ($\epsilon = 0.2$), we aggregated organ-level visitation counts from the CSF map across 100 training epochs. We normalized the visiting counts by the corresponding values obtained under uniform sampling, and normalized CSF patterns revealed organ-level differences in sampling preference. In particular, APS selectively concentrated sampling on anatomically

Fig. 3. Global and organ-level results of segmentation performance. **(A)** Dice score comparison of 21 organs for PET-only and PET-CT multimodal segmentation. Error bar indicates the standard deviation over 5-fold cross-validation. **(B)** Average dice score(y-axis) across 21 organs for U-Net and U-Net + PE trained with uniform sampling and adaptive patch sampling (APS), with the balancing parameter ϵ ranging from 0 to 0.9 (x-axis). **(C)** Organ-wise segmentation improvement achieved by APS relative to uniform sampling. **(D)** Stratification of 21 organs based on performance gains from the PE block(x-axis) and APS(y-axis). Dice scores are averaged across models trained with APS, ϵ ranging from 0.1 to 0.9; error bars indicate standard deviation across 5-fold cross-validation. Pearson(R_p) and Spearman(R_s) correlations are computed including and excluding the pancreas(light green), with statistical significance indicated(* $p < 0.05$, ** $p < 0.005$). **(E)** Validation loss and validation Dice score(y-axis) over the first 50 training epochs. Optimal configuration(U-Net: $\epsilon = 0.2$, U-Net + PE: $\epsilon = 0.3$) of APS (red) are compared with uniform sampling (black). Faster convergence is confirmed over 100 training epochs using Page's trend test ($p < 0.001$ for all model-metric combinations). Shaded regions represent standard deviation across 5-fold cross-validation.

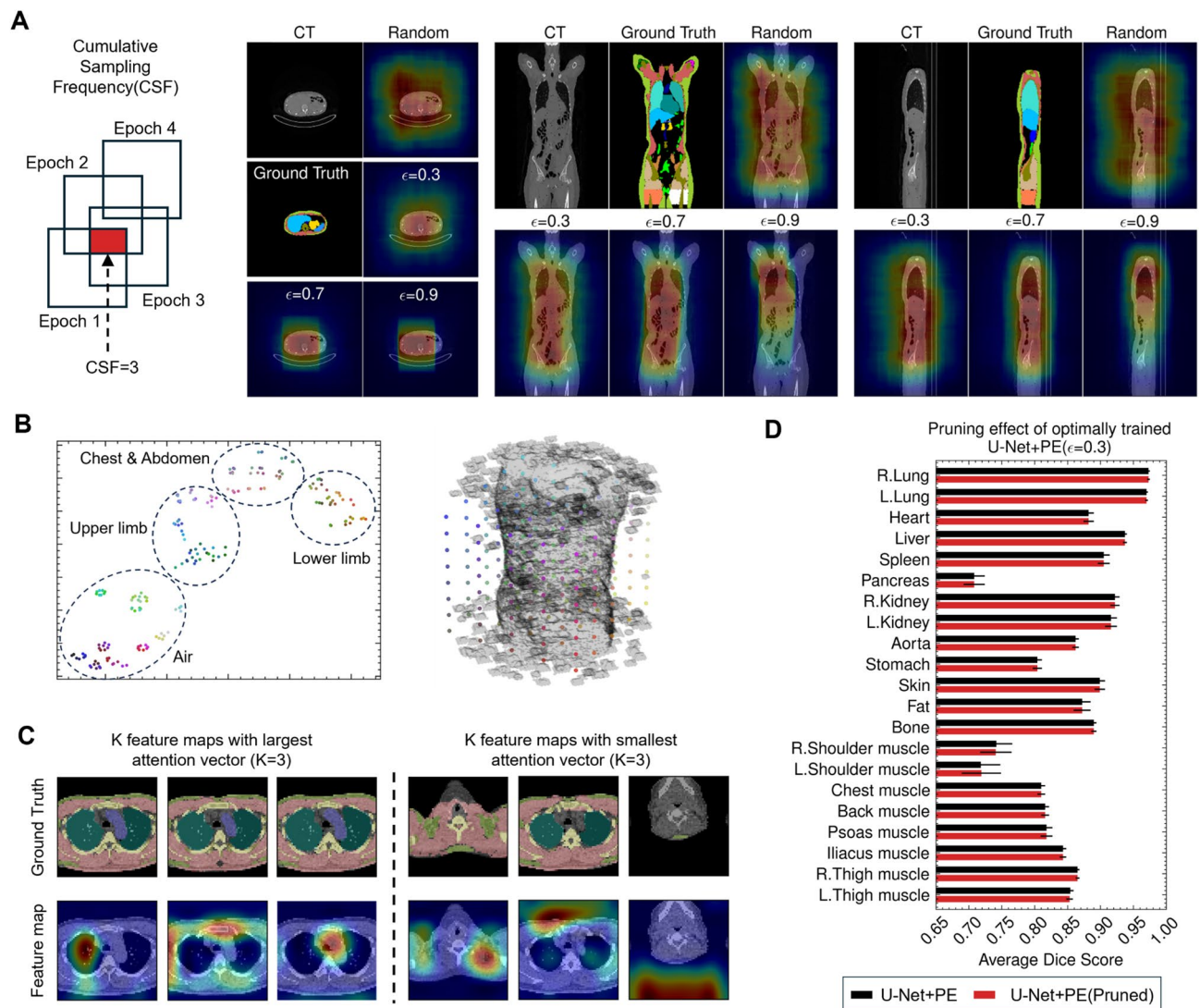


Fig. 4. Mechanistic interpretability of APS and the PE block. **(A)** Cumulative sampling frequency (CSF) for APS with $\epsilon=0.3, 0.7, 0.9$, shown in axial, coronal, and sagittal views and overlaid on CT images. **(B)** Two-dimensional t-SNE embedding of sampled patches across the whole body. Dot color (R, G, B) encodes the normalized spatial coordinates (ratio of patch coordinate to the size of the original image) of patch centers within the original image volume. **(C)** Feature map visualization for channels corresponding to the three highest and lowest organ attention values. Feature maps are upsampled to match the original image resolution and overlaid on CT images. **(D)** Effect of attention-guided feature map pruning on the optimally trained U-Net+PE model($\epsilon = 0.3$). Error bars indicate standard deviation across five cross-fold validation.

complex interfaces involving retroperitoneal organs on the right side of the body, including the right kidney and pancreas, as well as adjacent structures such as the liver (Supplementary Fig. 4).

Mechanistic interpretability of PE block

To examine *what* spatial context is encoded by the PE block, we evaluated whether it accurately infers anatomical information learned from organ existence patterns. We assessed organ existence prediction performance by comparing U-Net + PE with a baseline classifier network (CLS), a cascaded architecture comprising the U-Net encoder followed by the PE block encoder. The U-Net + PE model achieved an average F1-score of 0.908 in organ existence prediction (Supplementary Fig. 5), outperforming the baseline classifier and confirming that the PE block accurately captures organ context in each patch.

We then validated that organ attention vectors from the PE block encode meaningful spatial information by complementary analyses. Low-dimensional t-SNE projections of the organ attention vectors revealed a coherent anatomical organization: non-body patches formed distinct clusters, while body-region patches arranged in gradients consistent with their 3D spatial coordinates (Fig. 4B). Furthermore, feature map channels corresponding to the top $K = 3$ highest-valued organ attention vectors highlighted organ boundaries such as abdominal wall and lungs, whereas channels associated with the K lowest values lacked such anatomical correspondence (Fig. 4C). Pruning feature map channels with attention values below the threshold of 0.5 resulted in negligible accuracy loss, reinforcing the conclusion that the organ attention vector selectively amplifies channels relevant to segmentation while suppressing less informative ones (Fig. 4D). These interpretability analyses demonstrate that the PE block accurately predicts organ existence, produces anatomically coherent organ attention vectors, and the PE block is functionally valid, which is confirmed by causal intervention on channel attention values.

Discussion

This work presents a unified learning framework that rethinks patch-wise training as a dynamic process involving both adaptive data selection and context-aware representation learning. Patch-wise strategies are essential for large-scale dense prediction problems, yet static sampling and context-agnostic representations limit their efficiency and robustness. By explicitly addressing *where* computation is allocated through adaptive patch sampling (APS) and *what* contextual information is encoded through the patch encoding (PE) block, the proposed approach advances patch-based learning beyond conventional pipelines.

From an optimization perspective, APS reformulates patch sampling as an online, model-state-dependent importance sampling strategy, enabling efficient reallocation of training effort toward informative regions as optimization progresses. In parallel, the PE block complements APS by addressing a fundamental representational limitation of patch-wise learning: the loss of global context when operating on isolated patches. PE block mitigates the loss of global context inherent to patch-wise learning by conditioning feature representations on inferred contextual cues via channel-wise attention, without relying on explicit spatial coordinates or additional annotations. Together, these components improve convergence behavior and sample efficiency while remaining compatible with standard network architectures. A distinctive aspect of this work is its emphasis on analyzing learning dynamics. By explicitly modeling the sampling process and contextual conditioning, the framework enables mechanistic insight into how adaptive sampling and representation modulation influence optimization, complementing empirical performance evaluation.

APS showed interpretable sensitivity to its key hyperparameters: τ influenced the stability of early training dynamics, whereas ϵ governed the exploration–exploitation trade-off, with consistently strong performance around 0.2–0.3 across architectures. Although principled selection of these parameters remains an open problem, APS exhibited stable empirical behavior in practice. The current PE block was implemented only at the deepest bottleneck and depends on organ-existence supervision, limiting its present use to the fully supervised setting. Future work will investigate multi-resolution integration and more label-efficient supervision strategies.

Although demonstrated on whole-body PET-CT segmentation, the proposed framework is general and applicable to a broad class of large-scale dense prediction problems that rely on patch-based training. Future work may explore tighter theoretical characterization of adaptive sampling strategies and extensions to other forms of data selection and representation conditioning. In summary, this study contributes an efficient and interpretable learning strategy for adaptive patch-wise training, offering both practical benefits and algorithm-level insight for large-scale neural computation.

Data availability

Synapse Dataset can be found from: <https://www.synapse.org/Synapse:syn3193805>.

Received: 17 February 2026; Accepted: 24 April 2026

Published online: 15 May 2026

References

1. Xue, Z. et al. Multi-modal co-learning for liver lesion segmentation on PET-CT images. *IEEE Trans. Med. Imaging*. **40** (12), 3531–3542 (2021).
2. Gatidis, S. et al. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. *Sci. Data*. **9** (1), 601 (2022).
3. Gatidis, S. et al. Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. *Nat. Mach. Intell.* **6** (11), 1396–1405 (2024).
4. Sundar, L. K. S. et al. Fully automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. *J. Nucl. Med.* **63** (12), 1941–1948 (2022).
5. Klyuzhin, I. S. et al. PSMA-Hornet: Fully-automated, multi-target segmentation of healthy organs in PSMA PET/CT images. *Med. Phys.* **51** (2), 1203–1216 (2024).

6. Suganuma, Y. et al. Hybrid multiple-organ segmentation method using multiple U-Nets in PET/CT images. *Appl. Sci.* **13** (19), 10765 (2023).
7. Kim, H. et al. Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. *Sci. Rep.* **10** (1), 6204 (2020).
8. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods.* **18** (2), 203–211 (2021).
9. Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574–584). (2022).
10. Tang, Y. et al. 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* (pp. 20730–20740). (2022).
11. Moglia, A., Cavicchioli, M., Mainardi, L. & Cerveri, P. Deep learning for pancreas segmentation: a systematic review. *arXiv preprint arXiv:2407.16313*. (2024).
12. Tian, M., He, J., Yu, X., Cai, C. & Gao, Y. MCMC guided CNN training and segmentation for pancreas extraction. *IEEE Access.* **9**, 90539–90554 (2021).
13. Berger, L., Eoin, H., Cardoso, M. J. & Ourselin, S. An adaptive sampling scheme to efficiently train fully convolutional networks for semantic segmentation. In *Annual Conference on Medical Image Understanding and Analysis* (pp. 277–286). Cham: Springer International Publishing. (2018), July.
14. Wu, J., Tan, H., Liu, H., Liu, C. & Onofrey, J. A. An Adaptive Patch Sampling Scheme for Deep Learning Based PET Image Denoising. In *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)* (pp. 1–3). IEEE. (2022), November.
15. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Cham: Springer international publishing. (2015), October.
16. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention* (pp. 424–432). Cham: Springer International Publishing. (2016), October.
17. Zimmer, V. A., Gomez, A., Skelton, E., Wright, R., Wheeler, G., Deng, S., ... Schnabel, J. A. (2023). Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view. *Medical image analysis*, 83, 102639.
18. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). (2018).
19. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19). (2018).
20. Bastidas, A. A. & Tang, H. Channel attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 0–0). (2019).
21. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11534–11542). (2020).
22. Salaün, C., Huang, X., Georgiev, I., Mitra, N. J. & Singh, G. Online Importance Sampling for Stochastic Gradient Optimization. In *14th International Conference on Pattern Recognition Applications and Methods* (pp. 130–140). SciTePress. (2025).
23. Taylor, K. M. & Meyer, F. G. A random walk on image patches. *SIAM J. Imaging Sci.* **5** (2), 688–725 (2012).
24. Zhao, P. & Zhang, T. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning* (pp. 1–9). PMLR. (2015), June.
25. Needell, D., Srebro, N. & Ward, R. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Advances in neural information processing systems*, 27. (2014).
26. Alain, G., Lamb, A., Sankar, C., Courville, A. & Bengio, Y. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*. (2015).
27. Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7482–7491). (2018).
28. Yang, D., Li, Y. & Yu, J. Multi-task thyroid tumor segmentation based on the joint loss function. *Biomed. Signal Process. Control.* **79**, 104249 (2023).
29. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J. C., Pujol, S., ... Kikinis, R. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic resonance imaging*, 30(9), 1323–1341.
30. MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge. Synapse multi-organ segmentation dataset. (2015). <https://www.synapse.org/#!Synapse:syn3193805/wiki/217789> Accessed: 2023-07-10.
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
32. Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., ... Feng, A. (2022). Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.
33. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*. (2016).
34. Xu, J., Li, Z., Du, B., Zhang, M. & Liu, J. Reluplex made more practical: Leaky ReLU. In *2020 IEEE Symposium on Computers and communications (ISCC)* (pp. 1–7). IEEE. (2020), July.
35. Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. (2014).
36. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*. (2016).
37. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26** (3), 297–302 (1945).
38. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., ... Xie, W. (2025). Large-vocabulary segmentation for medical images with text prompts. *NPJ Digital Medicine*, 8(1), 566.
39. Derrac, J., García, S., Hui, S., Suganthan, P. N. & Herrera, F. Analyzing convergence performance of evolutionary algorithms: A statistical approach. *Inf. Sci.* **289**, 41–58 (2014).

Acknowledgements

This study was supported by a Severance Hospital Research fund for Clinical excellence (SHRC) (C-2023-0020) and by a student research bursary of Yonsei University, College of Medicine, and a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea. This research was supported by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00401794). The authors thank Mr. Maengkeun Oh for his data preparation.

Author contributions

H.P. initiated the project. J.P. conducted the methods and A.C. prepared the data. H.P. and J.P. wrote the main

manuscript text. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-51023-x>.

Correspondence and requests for materials should be addressed to H.-J.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026