

SAGE: Shape-Adapting Gated Experts for Adaptive Histopathology Image Segmentation

Gia Huy Thai*
University of Science, VNU-HCM
23120008@student.hcmus.edu.vn

Anh-Minh Phan
University of Technology, VNU-HCM
minh.phananh@hcmut.edu.vn

Thi-Ngoc-Truc Nguyen
Trivita AI
truc.nguyen@trivita.ai

Hoang-Nguyen Vu*
Trivita AI
nguyen.vu@trivita.ai

Quang-Thinh Ly
Michigan State University, USA
lythinh@msu.edu

Nhat Ho
The University of Texas at Austin
minhnhhat@utexas.edu

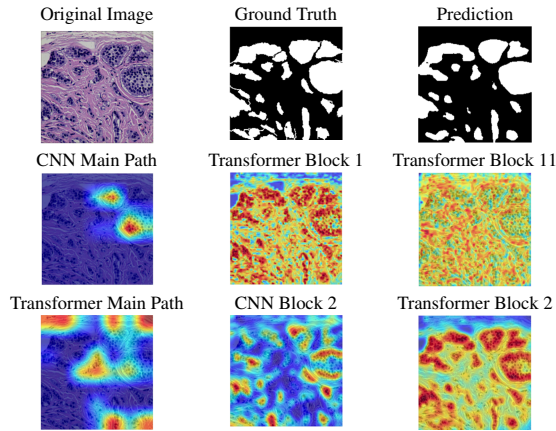


Figure 1. **Explainability visualization of dynamic expert routing in SAGE on the EBHI dataset.** From left to right, the first row shows the input patch, its ground-truth mask, and the predicted segmentation. Rows two and three report representative Grad-CAM heatmaps from the ConvNeXt and Transformer streams, including the main paths and selected expert blocks. Warmer regions indicate higher attribution, illustrating how SAGE routes computation to different experts to refine local boundaries while preserving global context under heterogeneous tissue appearances.

Abstract

The significant variability in cell size and shape continues to pose a major obstacle in computer-assisted cancer detection on gigapixel Whole Slide Images (WSIs), due to cellular heterogeneity. Current CNN-Transformer hybrids use static computation graphs with fixed routing.

*Equal Contribution.

This leads to extra computation and makes it harder to adapt to changes in input. We propose Shape-Adapting Gated Experts (SAGE), an input-adaptive framework that enables dynamic expert routing in heterogeneous visual networks. SAGE reconfigures static backbones into dynamically routed expert architectures via a dual-path design with hierarchical gating and a Shape-Adapting Hub (SA-Hub) that harmonizes feature representations across convolutional and transformer modules. Embodied as SAGE with ConvNeXt and Vision Transformer UNet (SAGE-ConvNeXt+ViT-UNet), our model achieves a Dice score of 95.23% on EBHI, DSC scores of 92.78% and 91.42% on GlaS Test A and Test B, respectively, and 91.26% DSC at the WSI level on DigestPath, while exhibiting robust generalization under distribution shifts by adaptively balancing local refinement and global context. SAGE establishes a scalable foundation for dynamic expert routing in visual networks, thereby facilitating flexible visual reasoning. Project page: <https://oxzygiahuy.github.io/sage/>.

1. Introduction

Computer-aided detection of malignant tissue in gigapixel WSIs is the basis of digital pathology. This makes it possible to quickly and accurately diagnose diseases. For quick diagnosis, classification, and treatment planning of colorectal cancer, it is very important to accurately describe the tumor’s morphology. Nonetheless, converting these visually intricate and diverse tissue architectures into computational comprehension continues to be exceedingly difficult. Convolutional Neural Networks (CNNs) [1] are great

at finding small local features like cell boundaries and textures. Vision Transformers (ViTs) [14], on the other hand, are a powerful way to model long-range spatial dependencies and global context. Nevertheless, substantial variability in tissue appearance, ranging from homogeneous normal tissues to complex and subtly textured malignant patterns, combined with the large resolution of WSIs, pushes current models beyond their representational and computational limits. Existing models, including U-Net variants and hybrid CNN-Transformer architectures, utilize a static computational graph. This makes all input segments go through the same processing, which is not a good way to do things because it over-processes simple areas and under-models complex ones. Also, the fact that CNN and Transformer blocks can only interact in one way means that you can't take advantage of each paradigm's strengths based on the characteristics of the input.

To address these limitations, we propose *Shape-Adapting Gated Experts* (SAGE), a dynamic, input-adaptive framework that converts a static backbone into a dual-path architecture. Each layer contains a main path that preserves the backbone transformation and an expert path that conditionally activates a subset of reused backbone blocks. A hierarchical router first estimates a group-level preference between shared and fine-grained expert groups, then applies top- K selection on prior-modulated logits to determine the active experts for each input. The two paths are fused adaptively, allowing the model to balance stability and input-specific refinement at run time. To enable interaction across heterogeneous experts (e.g., CNN and Transformer blocks), we introduce the *Shape-Adapting Hub* (SA-Hub), which aligns feature formats before and after expert execution. Although trained on patches, SAGE is deployed on full WSIs through sliding-window reconstruction, preserving compatibility with high-resolution pathology workflows.

To summarize, this work makes contributions as follows:

- We propose a dual-path formulation that transforms static backbones into dynamically routed architectures, enabling input-adaptive computation with parameter reuse.
- We design a hierarchical router with group-level gating and top- K selection over prior-modulated logits to balance shared and fine-grained specialization.
- We introduce SA-Hub, a lightweight shape-adaptation module that aligns CNN/Transformer feature formats for stable cross-expert communication.

2. Related Work

Medical Image Segmentation. Medical image segmentation is a core component of computational pathology because it enables quantitative analysis of cellular and tissue morphology. Earlier methods based on intensity thresholds, region growing, and contour evolution are sensitive

to noise and staining variation. Deep learning substantially improved robustness, starting from U-Net [25] and extending to stronger encoders such as ResNet [12], EfficientNet [31], ConvNeXt [17], and nested designs such as U-Net++ [39]. Recent studies emphasize data efficiency and transferability: foundation models such as MedSAM [19] and SAM-Med2D/3D [30] leverage large-scale pretraining, while semi-supervised methods such as C2GMatch [23] improve performance under limited annotations. However, severe domain heterogeneity in histopathology still makes robust cross-domain adaptation challenging.

Hybrid U-Net Architectures. Hybrid CNN-Transformer segmentation models seek to combine local detail modeling and long-range context. Representative architectures include TransUNet [5], Swin-UNet [2], and SegFormer [36]. Although these models improve global reasoning, they still follow static fusion and static computation graphs, which can be suboptimal for highly variable tissue patterns. State-space alternatives such as U-Mamba [20] and Swin U-Mamba [16] reduce complexity for large images via linear-time sequence modeling, but fine boundary precision can remain challenging in difficult regions. MoE-based segmentation models (e.g., MoE-NuSeg [35]) introduce conditional computation, yet most existing designs do not explicitly combine depth-wise adaptive routing with cross-architecture shape alignment.

Mixture of Experts. MoE provides scalable conditional computation by activating only a subset of experts per input [27]. Subsequent advances improve routing robustness and efficiency, including large-scale routing designs [8] and sigmoid-based gating [22, 37]. In vision, MoE has been explored for multimodal routing and multi-task adaptation [6, 21], and decoder-centric parameterization strategies [38]. Most prior approaches perform routing at token or spatial granularity, with limited depth-wise control over which layers are executed. MoLex [33] addresses this gap by treating layers as experts and routing across depth. SAGE builds on this direction and further introduces hierarchical group-aware routing together with shape-adaptive interaction between heterogeneous experts.

3. Method

Modern hybrid CNN-Transformer architectures generally depend on static computation graphs, applying the same sequence of operations to all inputs regardless of structural intricacy. While this design is stable and easy to implement, it limits adaptability toward heterogeneous visual patterns. To tackle this limitation, Shape-Adapting Gated Experts (SAGE) reparameterizes a fixed backbone into a two-path architecture with conditional expert routing, as illustrated in Figure 2. This framework preserves the original backbone pathway while introducing sparse expert selection for

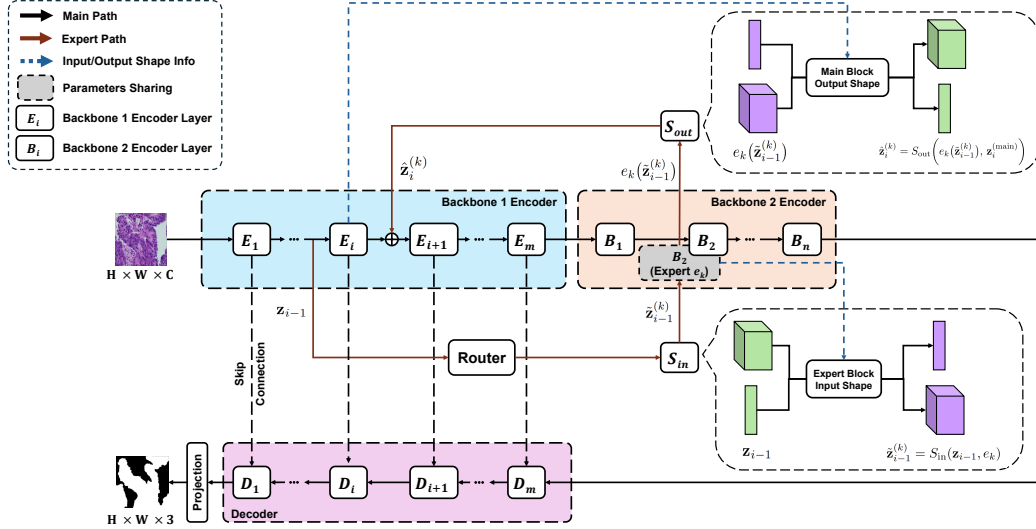


Figure 2. **Overview of the SAGE framework.** The main path (black arrows) keeps the original forward flow through Backbone 1 and Backbone 2, with multi-scale skip connections to the decoder. In parallel, the expert path (brown arrows) routes features from an intermediate layer to a sparse expert set (illustrated with Stage 2, Expert k) selected by the router. The Shape-Adapting Hub performs bidirectional format alignment via S_{in} and S_{out} so that cross-backbone expert execution is shape-compatible before fusion with the main branch. Blue arrows indicate input/output shape constraints used by the adapters, and dashed boxes denote parameter sharing (expert upcycling from pretrained backbone blocks).

feature optimization. The Sparse Mixture-of-Experts formulation is introduced in Section 3.1; the SAGE block and hierarchical routing are then detailed in Section 3.2 and 3.3. Finally, Section 3.4 presents the Shape-Adapting Hub for cross-architecture feature adaptation and integration.

3.1. Preliminaries: Sparse Mixture-of-Experts

Our SAGE framework is rooted in the Sparse Mixture-of-Experts (SMoE) formulation [27], which increases capacity by selectively activating only a subset of expert sub-networks, keeping computation roughly constant. A standard SMoE layer comprises a *router* and a set of experts $\{E_j\}_{j=1}^M$. Given an input x , the SMoE output is

$$y = \sum_{j \in \mathcal{K}} G(x)_j E_j(x), \quad (1)$$

where \mathcal{K} is the index set of selected experts. There are multiple choices for implementing $G(x)$, but a simple and performant option is to apply a softmax over the Top- K logits of a linear layer, $G(x) := \text{Softmax}(\text{TopK}(x \cdot W_g))$, so only the chosen experts are evaluated.

During training, a common issue known as *router collapse* arises when only a few experts dominate the routing. To promote balanced utilization, an auxiliary *load-balancing loss* encourages uniform token distribution

across experts:

$$\mathcal{L}_{\text{load-balancing}} = M \cdot \sum_{j=1}^M f_j P_j, \quad (2)$$

where M is the total number of experts, f_j denotes the proportion of tokens distributed to expert j and P_j is the proportion of the gating probability assigned to expert j .

Despite its scalability advantages, conventional SMoE primarily decides *which experts* to activate in a single routing stage. However, many complex tasks require not only expert selection but also adaptive coordination across heterogeneous computation types. Motivated by this limitation, our SAGE framework generalizes SMoE by introducing hierarchical routing and heterogeneous expert coordination, enabling the model to adaptively determine both *who* computes and *how* computation is performed for each input.

3.2. SAGE Block Architecture

At each backbone layer i (Figure 2), SAGE replaces a single deterministic transformation with a *dual-path* block that preserves the original computation while enabling learnable refinement. Given an input feature map \mathbf{z}_{i-1} , we compute a baseline feature via the original backbone layer and an enriched feature via sparsely activated layer experts, and then fuse them with a learnable gate.

Main path (backbone preservation). The main branch applies the original transformation $f_i(\cdot)$ to obtain a stable

baseline feature, $\mathbf{z}_i^{(\text{main})} = f_i(\mathbf{z}_{i-1})$. This pathway anchors optimization and retains the inductive bias and pretrained initialization of the CNN/Transformer backbone.

Expert path (conditional refinement). In parallel, the same input is routed to a sparse set of experts. Following MoLEx-style sparse upcycling [33], experts are the pre-trained backbone layers themselves and their parameters are reused rather than replicated; the router activates only a small subset per input. A hierarchical router (Section 3.3) performs top- K selection and yields expert weights $\{w_k\}_{k \in \mathcal{K}}$, while the expert-path feature is formed by weighted aggregation into $\mathbf{z}_i^{(\text{expert})}$. The exact construction of $\mathbf{z}_i^{(\text{expert})}$, including shape-adaptive translation and aggregation, is detailed in Section 3.4 and Equations 8–10.

Adaptive fusion. We gate between the baseline and expert-refined features using a learnable scalar α_i :

$$\mathbf{z}_i = \alpha_i \cdot \mathbf{z}_i^{(\text{main})} + (1 - \alpha_i) \cdot \mathbf{z}_i^{(\text{expert})}, \quad (3)$$

where $\alpha_i = \sigma(\theta_i)$ is computed from a learnable parameter θ_i . This formulation allows the model to dynamically balance stability and adaptability, favoring expert-driven refinement when beneficial while preserving the backbone’s inductive biases when necessary.

Expert pool composition. All expert paths draw from a global pool \mathcal{E} with a predefined number of fine-grained experts and shared experts. Both types are implemented as reused backbone layers, but they serve different objectives: fine-grained experts $\mathcal{E}_{\text{fine}}$ focus on depth-specific specialization, whereas shared experts $\mathcal{E}_{\text{shared}}$ encourage domain-generalizable computation.

3.3. Hierarchical Expert Routing

SAGE employs a two-level routing strategy to construct a sparse, input-dependent expert path while preserving the backbone computation. As illustrated in Figure 3, given a layer input \mathbf{z}_{i-1} , the router computes a group-level gate g_s , produces base expert logits via Semantic Affinity Routing (SAR), modulates these logits with the group prior, and then performs top- K selection on the modulated logits. This process jointly controls which experts are preferred (shared versus fine-grained) and which experts are finally executed, yielding the sparse expert computation used in Equation 3.

Group-Level Gating. A lightweight gating network G_s estimates the group-level preference toward shared experts. It takes a globally pooled representation $\bar{\mathbf{z}}_{i-1} \in \mathbb{R}^d$ and outputs a scalar gate $g_s \in (0, 1)$:

$$g_s = \sigma \left(\bar{\mathbf{z}}_{i-1} \mathbf{W}_{\text{gate}}^{(i)} + b_{\text{gate}}^{(i)} \right). \quad (4)$$

A high g_s favors shared experts, while a low g_s favors fine-grained experts; shared experts are still selected conditionally via the same top- K routing process.

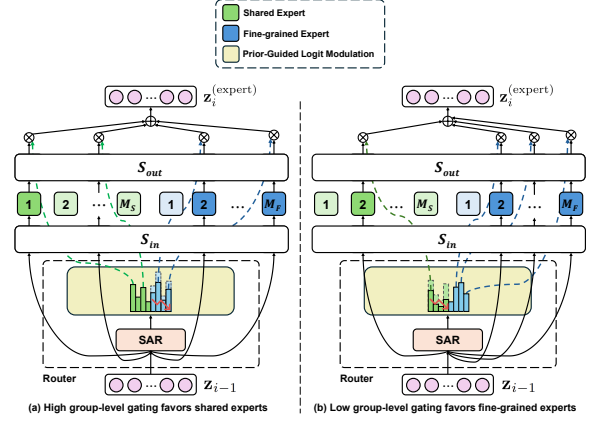


Figure 3. **Hierarchical routing with prior-guided logit modulation (Top- $K = 4$).** (a) For high g_s , shared-expert candidates (green) receive a weaker penalty than fine-grained candidates (blue), leading higher Top- K rank. (b) For low g_s , the preference reverses, favoring fine-grained experts for input-specific adaptation. Here, M_S and M_F denote the numbers of experts in the shared group and fine-grained group, with total pool size $M = M_S + M_F$; darker fills indicate activated experts. In both regimes, the router computes SAR logits, applies group-prior modulation, performs Top- K selection, and aggregates shape-adapted outputs via S_{in} and S_{out} into $\mathbf{z}_i^{(\text{expert})}$.

Semantic Affinity Routing (SAR). A primary router R_i computes base logits $\mathbf{L}_i \in \mathbb{R}^M$ over all experts $\mathcal{E} = \{E_1, \dots, E_M\}$:

$$\mathbf{L}_i = \frac{(\bar{\mathbf{z}}_{i-1} \mathbf{W}_Q^{(i)})(\mathbf{K}^{(i)})^\top}{\sqrt{d_k}} + \text{sp}(\bar{\mathbf{z}}_{i-1} \mathbf{W}_{\text{noise}}^{(i)}) \odot \boldsymbol{\epsilon}^{(i)}, \quad (5)$$

where $\text{sp}(\cdot)$ is the softplus function, $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$, $\mathbf{W}_Q^{(i)} \in \mathbb{R}^{d \times d_k}$ is a learnable query projection, $\mathbf{K}^{(i)} \in \mathbb{R}^{M \times d_k}$ is the expert-key matrix, and $\mathbf{W}_{\text{noise}}^{(i)} \in \mathbb{R}^{d \times M}$ controls input-adaptive noise magnitude. The first term captures semantic affinity, while the second term adds stochastic exploration to improve routing diversity and reduce expert over-specialization. The group-level preference induced by g_s is not applied within SAR; instead, it is introduced afterward via logit modulation.

Prior-Guided Logit Modulation. Our hierarchical gating couples *group-level* preference (shared vs. fine-grained experts) with *expert-level* selection. Given the base routing logits \mathbf{L}_i produced by SAR and the scalar shared gate g_s , we bias expert scores toward the preferred group before selecting the active experts, without forcing shared experts to remain always active. We define a binary mask \mathbf{m}_s , where $(\mathbf{m}_s)_j = 1$ if expert $j \in \mathcal{E}_{\text{shared}}$ and 0 otherwise. The modulated logits are obtained as:

$$\mathbf{L}'_i = \mathbf{L}_i + \mathbf{m}_s \log(g_s) + (\mathbf{1} - \mathbf{m}_s) \log(1 - g_s), \quad (6)$$

where $\mathbf{1} \in \mathbb{R}^M$ is the all-ones vector. This log-space prior

Algorithm 1 SAGE Training Algorithm (per mini-batch)

Require: Input batch \mathbf{X} with labels \mathbf{Y} ; model \mathcal{F} with T SAGE layers

Ensure: Total loss $\mathcal{L}_{\text{total}}$ for backpropagation

```
1:  $z_0 \leftarrow \text{Stem}(\mathbf{X})$ 
2:  $\mathcal{L}_{\text{load-balancing}} \leftarrow 0$ 
3: for  $i = 1$  to  $T$  do
4:    $z_i^{(\text{main})} \leftarrow f_i(z_{i-1})$ 
   Group-Level Gating and SAR
5:    $\bar{z}_{i-1} \leftarrow \text{GlobalPool}(z_{i-1})$ 
6:    $g_s \leftarrow \sigma(\bar{z}_{i-1} \mathbf{W}_{\text{gate}}^{(i)} + b_{\text{gate}}^{(i)})$ 
7:    $\mathbf{L}_i \leftarrow \text{SAR}(\bar{z}_{i-1})$ 
   Prior-Guided Logit Modulation
8:    $g_s \leftarrow \text{clip}(g_s, \epsilon, 1 - \epsilon)$ 
9:    $\mathbf{L}'_i \leftarrow \mathbf{L}_i + \mathbf{m}_s \log(g_s) + (\mathbf{1} - \mathbf{m}_s) \log(1 - g_s)$ 
   Top- $K$  Sigmoid Gating and Expert Execution
10:   $\mathcal{K} \leftarrow \text{TopKIndices}(\mathbf{L}'_i, K)$ 
11:   $\mathbf{w}[\mathcal{K}] \leftarrow \sigma(\mathbf{L}'_i[\mathcal{K}]); \mathbf{w}[\bar{\mathcal{K}}] \leftarrow 0$ 
12:   $z_i^{(\text{expert})} \leftarrow 0$ 
13:  for  $k \in \mathcal{K}$  do
14:     $\tilde{z}_{i-1}^{(k)} \leftarrow S_{\text{in}}(z_{i-1}, e_k)$ 
15:     $\hat{z}_i^{(k)} \leftarrow S_{\text{out}}(e_k(\tilde{z}_{i-1}^{(k)}), z_i^{(\text{main})})$ 
16:     $z_i^{(\text{expert})} \leftarrow z_i^{(\text{expert})} + \mathbf{w}_k \hat{z}_i^{(k)}$ 
17:  end for
18:   $z_i \leftarrow \alpha_i z_i^{(\text{main})} + (1 - \alpha_i) z_i^{(\text{expert})}$ 
19:   $\mathcal{L}_{\text{load-balancing}} \leftarrow \mathcal{L}_{\text{load-balancing}} + M \sum_{j=1}^M f_j^{(i)} P_j^{(i)}$ 
20: end for
21:  $\mathbf{P} \leftarrow \text{Decoder}(z_T)$ 
22:  $\mathcal{L}_{\text{task}} \leftarrow \lambda_{\text{ce}} \mathcal{L}_{\text{CE}}(\mathbf{P}, \mathbf{Y}) + \lambda_{\text{dice}} \mathcal{L}_{\text{Dice}}(\mathbf{P}, \mathbf{Y})$ 
23:  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}} + \lambda_{\text{lb}} \mathcal{L}_{\text{load-balancing}}$ 
24: return  $\mathcal{L}_{\text{total}}$ 
```

raises shared-expert logits when g_s is high and fine-grained logits when g_s is low. For numerical stability, we clip g_s to $[\epsilon, 1 - \epsilon]$, then select $\mathcal{K} = \text{TopKIndices}(\mathbf{L}'_i, K)$. Unlike softmax gating, we use independent sigmoid gates on the selected experts:

$$w_j = \sigma((\mathbf{L}'_i)_j) \cdot \mathbf{1}[j \in \mathcal{K}], \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid function; $\mathbf{w} = \{w_j\}_{j=1}^M$. For $k \in \mathcal{K}$, w_k denotes the gate value at expert index k . This design permits independent multi-expert activation without simplex normalization across selected experts.

Overall, the integration of logit modulation with top- K sigmoid gating induces a sparse, input-adaptive computation graph that balances efficiency with expert specialization across heterogeneous feature distributions.

3.4. Cross-Architecture Adaptation and Fusion

Executing heterogeneous experts requires resolving representation mismatches between convolutional feature maps (B, C, H, W) and token sequences (B, N, D) . We address

this with the Shape-Adapting Hub, a lightweight learnable module that performs bidirectional format conversion with explicit shape alignment.

Shape-Adapting Hub (SA-Hub). SA-Hub consists of an input adapter S_{in} and an output adapter S_{out} . For each selected expert e_k , we first infer the representation type of the source feature and target expert, then apply a format-aware transformation:

$$\tilde{\mathbf{z}}_{i-1}^{(k)} = S_{\text{in}}(\mathbf{z}_{i-1}, e_k), \quad (8)$$

with

$$S_{\text{in}}(\mathbf{x}, e_k) = \begin{cases} \mathcal{A}_{C \rightarrow D}(\mathcal{F}_{\text{tokenize}}(\mathbf{x})), & \tau(\mathbf{x}) = \text{CNN}, \\ & \tau(e_k) = \text{Trans.}, \\ \mathcal{F}_{\text{detokenize}}(\mathcal{A}_{D \rightarrow C}(\mathbf{x})), & \tau(\mathbf{x}) = \text{Trans.}, \\ & \tau(e_k) = \text{CNN}, \\ \mathcal{A}(\mathbf{x}), & \tau(\mathbf{x}) = \tau(e_k), \end{cases}$$

where $\tau(\cdot)$ denotes the representation format type (e.g., CNN feature map or Transformer token sequence). $\mathcal{F}_{\text{tokenize}}$ maps spatial feature maps to token sequences, and $\mathcal{F}_{\text{detokenize}}$ reconstructs spatial layouts (with interpolation if needed). $\mathcal{A}_{C \rightarrow D}$ and $\mathcal{A}_{D \rightarrow C}$ denote projections between channel and token-embedding spaces, while $\mathcal{A}(\cdot)$ is used when source and target formats already match. After expert computation, the output adapter maps back to the main-path format with target-shape constraints:

$$\hat{\mathbf{z}}_i^{(k)} = S_{\text{out}}\left(e_k(\tilde{\mathbf{z}}_{i-1}^{(k)}), \mathbf{z}_i^{(\text{main})}\right), \quad (9)$$

where S_{out} applies the inverse format transformation and enforces the spatial resolution and channel dimensionality of $\mathbf{z}_i^{(\text{main})}$. Finally, the adapted expert outputs are combined by gating-weighted aggregation:

$$\mathbf{z}_i^{(\text{expert})} = \sum_{k \in \mathcal{K}} w_k \cdot \hat{\mathbf{z}}_i^{(k)}. \quad (10)$$

This design allows CNN and Transformer experts to be mixed within a routing layer without imposing a shared native tensor format. In implementation, SA-Hub maintains a registry of pre-initialized adapters for common dimensional conversions to reduce runtime overhead while preserving end-to-end differentiability. The complete execution flow of SAGE, including hierarchical gating, shape adaptation, and loss computation, is summarized in Algorithm 1.

4. Experiments

4.1. Datasets and Evaluation Metrics

We rigorously evaluated the SAGE framework using three established public benchmarks for colorectal histopathology segmentation: *EBHI*, *GlaS*, and *DigestPath*.

Table 1. **Baseline comparison on EBHI and GlaS.** SAGE-ConvNeXt+ViT-UNet consistently outperforms strong baseline backbones across all reported metrics. All scores are percentages (%). **Best** and **Second** indicate the best and second-best performance, respectively.

Model	EBHI (Adenocarcinoma)					GlaS (Test A)					GlaS (Test B)						
	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	O-DSC ↑	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	O-DSC ↑
ResNet101-UNet [11, 26]	91.93	89.11	94.24	46.15	55.58	87.52	77.75	87.49	32.37	60.11	41.68	85.72	78.46	87.93	26.10	56.69	40.11
ResNet152-UNet [11, 26]	91.64	88.69	94.00	45.93	54.89	87.60	77.76	87.60	33.08	55.79	33.81	84.28	76.73	86.83	26.14	49.10	27.69
EfficientNet-B7-UNet [26, 32]	91.82	89.00	94.18	46.37	54.45	87.92	78.77	88.12	33.57	53.26	22.99	84.97	77.14	87.09	30.26	51.25	24.33
ConvNeXt-UNet [17, 26]	91.87	89.07	94.22	46.01	54.59	91.91	85.03	91.91	23.39	68.03	54.31	88.22	81.68	89.91	25.85	59.37	49.17
U-Net++ (ResNet 101) [39]	91.94	89.13	94.25	45.51	56.00	89.01	80.40	89.13	30.21	61.14	29.21	86.23	79.30	88.46	26.34	53.08	19.38
UMamba [20]	91.51	88.55	93.93	49.79	51.66	83.64	72.17	83.83	34.31	53.14	36.40	84.46	77.22	87.15	26.23	54.89	45.00
Swin UMamba [16]	92.01	89.18	94.28	48.35	55.23	86.57	76.09	86.42	34.93	60.59	35.46	84.00	75.00	85.72	26.76	58.29	38.37
Swin U-Net [2]	91.97	89.10	94.24	45.73	53.06	90.53	82.72	90.54	26.89	60.20	45.20	86.82	79.52	88.59	24.82	54.70	43.09
ResNet34-UNet [11, 26]	91.72	88.63	93.88	47.41	54.67	88.91	81.72	89.78	26.48	63.84	55.26	87.46	80.68	88.92	24.62	57.38	49.84
SAGE-ResNet34-UNet (Ours)	92.36	89.24	94.28	46.05	56.18	90.37	83.34	91.06	23.39	69.27	61.18	89.28	82.11	90.41	21.27	63.45	56.73
ConvNeXt+ViT-UNet	92.49	83.60	90.94	46.02	55.60	91.88	84.91	91.80	24.25	75.21	65.09	89.96	81.80	89.85	22.32	67.57	59.93
SAGE-ConvNeXt+ViT-UNet (Ours)	94.03	90.90	95.23	45.20	58.10	92.96	86.62	92.78	19.85	77.91	73.49	91.55	84.56	91.42	17.94	70.23	66.67

Table 2. **SOTA comparison on EBHI and GlaS.** SAGE-ConvNeXt+ViT-UNet achieves the strongest overall performance across metrics, indicating robust generalization across diverse tissue morphologies. All metrics are reported in percentages (%). **Best** and **Second** indicate the best and second-best performance, respectively.

Model	EBHI (Adenocarcinoma)					GlaS (Test A)					GlaS (Test B)						
	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	O-DSC ↑	Acc ↑	IoU ↑	DSC ↑	HD95 ↓	B-F1 ↑	O-DSC ↑
SelfReg-UNet [41]	91.53	88.58	93.95	45.66	53.83	89.36	80.34	89.10	28.57	65.82	47.17	86.21	77.93	87.60	26.33	59.95	40.16
Attention U-Net [24]	92.11	89.28	94.34	46.43	57.64	90.53	82.43	90.37	65.60	45.95	59.03	87.57	80.29	89.07	56.51	38.87	54.77
ConvUNet [10]	90.73	87.66	93.43	47.93	49.86	87.68	77.61	87.39	28.70	62.80	45.60	88.05	81.16	89.60	20.46	60.58	42.96
UCTransNet [34]	91.95	89.07	94.22	46.67	57.22	87.07	76.47	86.66	29.23	62.97	49.38	85.45	77.14	87.10	26.66	58.51	58.00
TransAttUNet [3]	91.40	88.52	93.91	50.02	51.40	91.18	83.69	91.12	24.51	72.13	66.65	89.87	83.78	91.18	22.41	64.27	60.86
SegFormer [36]	92.62	89.93	94.70	42.43	55.16	91.09	83.31	90.89	20.72	61.65	53.72	87.47	80.45	89.16	23.60	55.95	51.54
EViT-UNet [15]	92.80	90.23	94.86	45.30	59.99	92.70	86.15	92.56	21.82	76.54	73.26	89.61	83.62	91.08	21.24	65.62	63.24
CAC-UNet [40]	91.32	88.40	93.84	51.35	53.42	88.01	78.34	87.85	31.70	64.29	53.81	85.69	77.52	87.33	27.87	58.11	46.02
TransUNet [4]	91.46	88.38	93.83	45.76	53.96	91.26	83.72	91.14	23.46	63.62	67.33	87.30	79.98	88.87	22.77	55.63	63.56
SAGE-ConvNeXt+ViT-UNet (Ours)	94.03	90.90	95.23	45.20	58.10	92.96	86.62	92.78	19.85	77.91	73.49	91.55	84.56	91.42	17.94	70.23	66.67

EBHI Dataset. The Extended Biopsy Histopathological Image (EBHI) dataset [28] contains 5,170 H&E-stained biopsy samples, each classified into one of six histological subtypes. We focused on the clinically significant *Adenocarcinoma* subset, and selected 795 images for our experiments. This dataset evaluates SAGE on heterogeneous yet domain-consistent tissue patterns.

GlaS Dataset. The Gland Segmentation (GlaS) dataset [29] was introduced in the MICCAI 2015 Gland Segmentation Challenge. It consists of 165 H&E-stained histology images at a resolution of 522×775 , each annotated for glandular structures. The official split includes 85 images for *Train*, 60 for *Test A*, and 20 for *Test B*. This dataset is widely used to assess a model’s ability to capture gland morphology and boundary precision.

DigestPath Dataset. The DigestPath dataset [7] was introduced in the DigestPath 2019 Challenge and contains 660 gigapixel whole-slide images (WSIs) from colonoscopy specimens. To facilitate efficient training, we devised a preprocessing pipeline applied to all WSIs. Each WSI was partitioned into overlapping tiles of size 1536×1536 with stride

512. A patch P was retained only when both conditions were satisfied on its grayscale intensities:

$$P \text{ is retained iff } \begin{cases} \sigma(P) \geq 10, \\ \mu(P) \leq 230, \end{cases}$$

where $\sigma(P)$ and $\mu(P)$ denote the standard deviation and mean intensity of P , respectively. This preprocessing yielded approximately 40,000 patches across all WSIs, filtering out low-information background while preserving diagnostically relevant tissue regions.

Evaluation Metrics. We assessed segmentation performance using a comprehensive set of metrics: pixel-wise Accuracy (ACC), Intersection over Union (IoU), Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Boundary F1 (B-F1). For GlaS, given its focus on object-level segmentation rather than semantic segmentation, we additionally reported the Object DSC score (O-DSC). ACC, IoU, DSC, B-F1, and O-DSC are reported in percentage (%), while HD95 is reported in pixels. For EBHI and GlaS, metrics are computed per image and averaged over the held-out test set. For DigestPath, we report both

patch-level and WSI-level results. To reconstruct WSI-level predictions, overlapping patch logits are mapped back to their original slide coordinates and merged by averaging in overlap regions, followed by a threshold of 0.5 to obtain the final binary mask; this reconstruction uses the same tiling stride (512) as preprocessing.

4.2. Implementation Details

SAGE-ConvNeXt+ViT-UNet Configuration. The SAGE module was integrated into the hybrid encoder combining ConvNeXt [17] and an ImageNet-pretrained ViT [9]. The MoE module included 4 shared experts and 16 non-shared experts, with the top 4 experts dynamically selected during routing. We trained SAGE-ConvNeXt+ViT-UNet on two NVIDIA H100 GPUs (80GB VRAM each) with a global batch size of 64 (32 per GPU). Training proceeded in two stages using the AdamW optimizer [18] and the hybrid loss $\mathcal{L}_{\text{total}}$ in Algorithm 1, weighted as $\lambda_{\text{ce}} = 1$, $\lambda_{\text{dice}} = 1.5$, $\lambda_{\text{fb}} = 1$. In the first stage, all parameters were optimized with a uniform learning rate of 1×10^{-5} . In the second stage, *discriminative fine-tuning* [13] was applied with 1×10^{-5} for non-shared experts, routers, and the decoder, and 5×10^{-5} for shared experts. Unless noted otherwise, all experiments used random seed 42, and other hyperparameters followed default settings.

Comparison Protocol. To benchmark SAGE-ConvNeXt+ViT-UNet, we compared its performance in two settings: *Baseline Comparison* and *State-of-the-Art (SOTA) Comparison*. For GlaS and EBHI, we report results under both settings, with all input images resized to 224×224 . In contrast, DigestPath was evaluated only under the SOTA setting, using a higher resolution of 512×512 to preserve critical tissue-level detail. We adhered to the official training setups for all methods (e.g., optimizer, learning rate), except for a standardized batch size of 64, which ensured efficient multi-GPU utilization.

Evaluation and Model Selection. Dataset splitting strategies varied by dataset structure. For GlaS, we split the official training data into 80% for training and 20% for validation, using the official Test A and Test B sets for final evaluation. For EBHI, we employed a random split of 70% training, 15% validation, and 15% testing at the image level. For DigestPath, we performed a stratified split of 70%/15%/15% based on the raw positive and negative WSIs. The patch extraction pipeline was subsequently applied to the WSIs within each split. For all datasets, the model checkpoint with the highest Dice score on the validation set was selected for final evaluation on the test set.

4.3. Quantitative Comparison

This section provides a quantitative comparison of SAGE-ConvNeXt+ViT-UNet against strong baselines and SOTA

methods. Baseline results are in Table 1, while SOTA comparisons are in Table 2 and Table 3.

Table 3. **SOTA comparison on DigestPath at Patch and WSI levels.** SAGE-ConvNeXt+ViT-UNet delivers the best overall results under both evaluation protocols. All metrics are reported in percentages (%). **Best** and **Second** indicate the best and second-best performance among all models, respectively.

Model	DigestPath									
	Patch					WSI				
	Acc \uparrow	IoU \uparrow	DSC \uparrow	HD95 \downarrow	B-F1 \uparrow	Acc \uparrow	IoU \uparrow	DSC \uparrow	HD95 \downarrow	B-F1 \uparrow
SelfReg-UNet [41]	96.28	80.05	83.32	131.20	71.61	97.77	78.06	83.04	417.45	60.09
Attention U-Net [24]	96.81	80.47	83.97	123.71	72.29	98.03	80.30	85.36	416.23	61.31
ConvUNetXt [10]	96.84	77.71	81.22	123.98	69.40	98.03	74.20	79.25	576.06	55.37
UCTransNet [34]	96.63	81.54	84.89	121.27	73.81	97.95	84.44	89.44	427.20	66.35
TransAttUNet [3]	96.86	80.37	83.69	126.87	71.37	98.02	75.53	80.45	454.48	56.00
SegFormer [36]	96.99	84.54	87.66	115.75	76.54	97.97	83.51	90.56	393.11	67.15
ViTUNet [15]	96.60	83.38	86.37	120.47	76.38	98.02	83.16	87.78	364.86	65.97
CAC-UNet [40]	96.81	81.74	84.99	61.10	78.78	97.96	83.63	88.66	528.06	65.03
TransUNet [4]	96.79	83.19	86.50	119.96	74.40	97.92	82.25	87.28	419.77	63.18
SAGE-ConvNeXt+ViT-UNet	96.80	89.51	91.96	132.88	75.91	97.95	83.46	88.69	482.08	65.16
SAGE-ConvNeXt+ViT-UNet (Ours)	97.69	90.21	92.66	60.67	79.48	98.73	86.21	91.26	362.31	67.85

Baseline Comparison. Table 1 demonstrates that SAGE enhances both pure-CNN and hybrid backbones, rather than benefiting a single architecture. In a direct plug-in comparisons, SAGE-ResNet34-UNet improves over ResNet34-UNet on EBHI and on both GlaS splits. The gains are especially clear on gland-level structure quality, where O-DSC increases by about +5.9%/+6.9% on Test A/B. A similar trend is observed in the hybrid pair: SAGE-ConvNeXt+ViT-UNet also outperforms its non-SAGE counterpart on EBHI, with strong gains in IoU and DSC (approximately +7.30% IoU and +4.29% DSC). Overall, these paired comparisons indicate that the performance gain is mainly attributable to dynamic routing, not to a particular encoder design.

SOTA Comparison. Tables 2 and 3 present a comparison between the proposed model and recent SOTA methods. SAGE-ConvNeXt+ViT-UNet achieves the highest overall performance across EBHI, GlaS, and DigestPath, with the most significant margins observed in more challenging settings. On EBHI and GlaS, it surpasses strong transformer-based competitors while maintaining the highest DSC values. On DigestPath, it ranks first at both patch and WSI levels, with a notably larger gain at WSI level (+2.57% DSC) compared to the patch level (+0.70% DSC). This pattern aligns with the objective of enhancing robustness to morphological variation and domain shift.

4.4. Efficiency and Model Complexity

Table 4 highlights two key points. First, despite adding dual-path routing and SA-Hub adapters, SAGE keeps parameter growth modest ($543.71\text{M} \rightarrow 573.51\text{M}$, +5.5%) because expert blocks are reused from the backbone. Second, computation scales with Top- K : FLOPs increase from 63.77G (baseline) to 99.51G ($K = 1$), 130.47G ($K = 2$), and 486.81G ($K = 4$), reflecting the expected accuracy-efficiency trade-off of sparse expert activation.

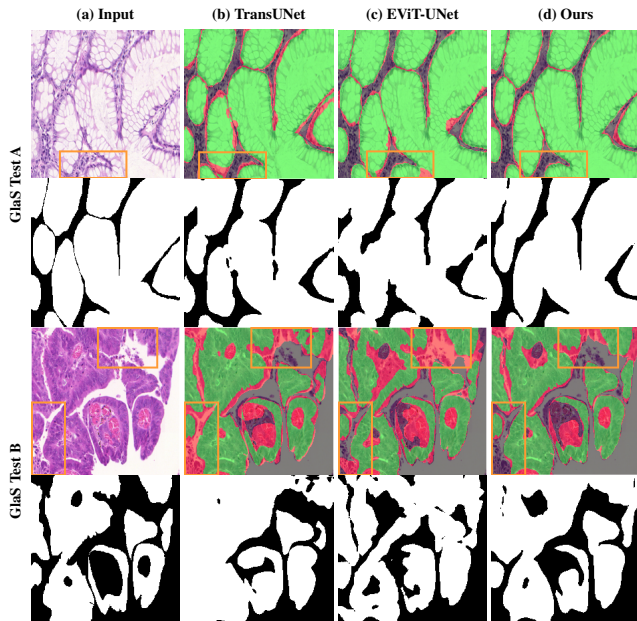


Figure 4. **Qualitative comparison on GlaS test samples.** Each column shows (a) the input image with ground-truth annotation, (b) TransUNet, (c) EViT-UNet, and (d) our proposed SAGE-ConvNeXt+ViT-UNet. The top row presents a typical gland structure (GlaS Test A), while the bottom row depicts a challenging case with irregular morphology (GlaS Test B). Green areas denote correct predictions, and red areas denote errors.

Table 4. **Model complexity across varying Top- K settings.** Parameter count and FLOPs per image for the baseline and SAGE.

Model	Top- K	Params (M)	GFLOPs/image
ConvNeXt+ViT-UNet (Baseline)	–	543.71	63.77
SAGE-ConvNeXt+ViT-UNet	1	573.51	99.51
SAGE-ConvNeXt+ViT-UNet	2	573.51	130.47
SAGE-ConvNeXt+ViT-UNet	4	573.51	486.81

4.5. Qualitative Results

Figure 4 provides qualitative evidence supporting the improvements reported in Table 1 and Table 2. On GlaS Test A, SAGE-ConvNeXt+ViT-UNet produces cleaner gland boundaries and fewer false positives compared to TransUNet and EViT-UNet. The performance gap increases on the more challenging Test B sample, where competing methods exhibit pronounced stromal over-segmentation and occasional gland merging, whereas SAGE maintains gland separation and boundary fidelity. These visual trends align with the quantitative gains and indicate enhanced robustness under domain shift.

To further assess deployment behavior, Figure 5 provides WSI-level qualitative comparisons on DigestPath. In WSI Sample 1, which features branching glandular structures, SAGE-ConvNeXt+ViT-UNet more accurately follows thin lesion boundaries and produces fewer false positives than

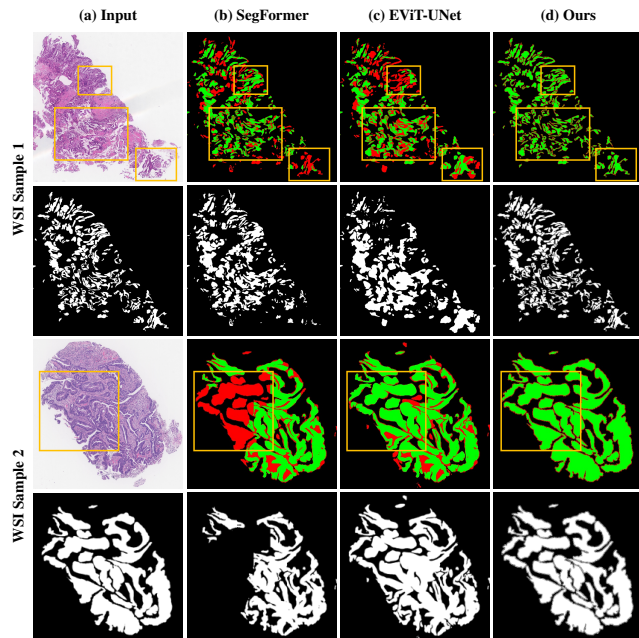


Figure 5. **Qualitative comparison on Digestpath WSI test samples.** Each column displays vertically stacked pairs of (a) the input image and ground-truth annotation, followed by the error overlays and binary predictions for (b) SegFormer, (c) EViT-UNet, and (d) our proposed SAGE-ConvNeXt+ViT-UNet. WSI Sample 1 (top rows) presents branching tissue structures requiring fine local boundary delineation, while WSI Sample 2 (bottom rows) depicts a dense, uniform tissue mass prone to patch-level attention collapse in pure transformer architectures. Green areas denote correct predictions, and red areas denote errors.

SegFormer and EViT-UNet. In the denser WSI Sample 2, competing models display spillover into background tissue and partial region collapse, while SAGE preserves contiguous lesion topology with cleaner boundaries. These WSI-level findings are consistent with the quantitative improvements reported in Table 3, indicating that dynamic routing enhances stability beyond patch-level predictions.

5. Conclusion

We introduced SAGE (Shape-Adapting Gated Experts), a dynamic and backbone-agnostic framework for histopathology image segmentation that replaces static CNN-Transformer fusion with hierarchical expert routing and a Shape-Adapting Hub to adapt computation to input morphology; across EBHI, GlaS, and DigestPath, SAGE consistently outperforms backbone-matched baselines and recent methods at both patch and WSI levels, while offering interpretable routing behavior that clarifies expert specialization, and future work will extend evaluation to broader clinical settings and additional dense prediction tasks.

Acknowledgements

We gratefully acknowledge The University of Texas at Austin for supporting this research, and Trivita AI and AI VIET NAM for providing the GPU computing resources essential to this work.

References

- [1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6, 2017. 1
- [2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2, 6
- [3] Bingzhi Chen, Yishu Liu, Zheng Zhang, Guangming Lu, and Adams Wai Kin Kong. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):55–68, 2023. 6, 7
- [4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6, 7
- [5] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 2
- [6] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17300–17311, 2023. 2
- [7] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, Hongmei Yi, Yan Guo, Zhe Wang, Ling Chen, Li Zhang, Xianying He, Xiaofan Zhang, Ke Mei, Chuang Zhu, Weizeng Lu, Linlin Shen, Jun Shi, Jun Li, Sreehari S, Ganapathy Krishnamurthi, Jiangcheng Yang, Tiancheng Lin, Qingyu Song, Xuechen Liu, Simon Graham, Raja Muhammad Saad Bashir, Canqian Yang, Shaofei Qin, Xinmei Tian, Baocai Yin, Jie Zhao, Dimitris N. Metaxas, Hongsheng Li, Chaofu Wang, and Shaoting Zhang. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022. 6
- [8] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseek-moe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. 2
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 7
- [10] Zhimeng Han, Muwei Jian, and Gai-Ge Wang. Convunet: An efficient convolution neural network for medical image segmentation. *Knowledge-based systems*, 253:109512, 2022. 6, 7
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2
- [13] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. 7
- [14] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2
- [15] Xin Li, Wenhui Zhu, Xuanzhao Dong, Oana M Dumitrescu, and Yalin Wang. Evit-unet: U-net like efficient vision transformer for medical image segmentation on mobile and edge devices. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. 6, 7
- [16] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024. 2, 6
- [17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. 2, 6, 7
- [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 7
- [19] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), 2024. 2
- [20] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2, 6
- [21] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156, 2016. 2
- [22] Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Sigmoid gating is more sample efficient than softmax gating in mixture of experts, 2024. 2
- [23] Thi-Ngoc-Truc Nguyen, Xuan-Hong Ong, Hoang-Thien-Nguyen, Van-Hung Bui, Hoang-Nguyen Vu, Thanh Minh Nguyen, Quan Nguyen, and Huu-Hung Nguyen. C2gmatch:

- Leveraging dual-view cross-guidance and co-guidance framework for semi-supervised cell segmentation. In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 1–6, 2025. [2](#)
- [24] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. [6](#), [7](#)
- [25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, Cham, 2015. [2](#)
- [26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. [6](#)
- [27] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. [2](#), [3](#)
- [28] Liyu Shi, Xiaoyan Li, Weiming Hu, Haoyuan Chen, Jing Chen, Zizhen Fan, Minghe Gao, Yujie Jing, Guotao Lu, Deguo Ma, Zhiyu Ma, Qingtao Meng, Dechao Tang, Hongzan Sun, Marcin Grzegorzec, Shouliang Qi, Yueyang Teng, and Chen Li. Ebhi-seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Frontiers in Medicine*, Volume 10 - 2023, 2023. [6](#)
- [29] Korsuk Sirinukunwattana, Josien P. W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R. J. Snead, and Nasir M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest, 2016. [6](#)
- [30] Jiahao Sun, Kai Chen, Zhiwei He, et al. Medical image analysis using improved sam-med2d: segmentation and classification perspectives. *BMC Medical Imaging*, 24:241, 2024. [2](#)
- [31] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. [2](#)
- [32] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. [6](#)
- [33] Rachel S. Y. Teo and Tan M. Nguyen. Molex: Mixture of layer experts for finetuning with sparse upcycling, 2025. [2](#), [4](#)
- [34] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Utransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2441–2449, 2022. [6](#), [7](#)
- [35] Xuening Wu, Yiqing Shen, Qing Zhao, Yanlan Kang, and Wenqiang Zhang. Moe-nuseg: Enhancing nuclei segmentation in histology images with a two-stage mixture of experts network. *Alexandria Engineering Journal*, 110:557–566, 2025. [2](#)
- [36] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. [2](#), [6](#), [7](#)
- [37] Fanqi Yan, Huy Nguyen, Pedram Akbarian, Nhat Ho, and Alessandro Rinaldo. Sigmoid self-attention is better than softmax self-attention: A mixture-of-experts perspective. *arXiv preprint arXiv: 2502.00281*, 2025. [2](#)
- [38] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture of low-rank experts, 2024. [2](#)
- [39] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer. Epub 2018 Sep 20. [2](#), [6](#)
- [40] Chuang Zhu, Ke Mei, Ting Peng, Yihao Luo, Jun Liu, Ying Wang, and Mulan Jin. Multi-level colonoscopy malignant tissue detection with adversarial cac-unet. *Neurocomputing*, 438:165–183, 2021. [6](#), [7](#)
- [41] Wenhui Zhu, Xiwen Chen, Peijie Qiu, Mohammad Farazi, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. Selfreg-unet: Self-regularized unet for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 601–611. Springer, 2024. [6](#), [7](#)