Expert Branches: Module Diversity for Stronger Feature Learning in Laparoscopic Segmentation

Lin Guo[1] LIN.GUO@SLU.EDU
Chiara Camerota[1] CHIARA.CAMEROTA@SLU.EDU
Mohammad Mahmoud[2] MMAHMOUD96@SIUMED.EDU
[2] *School of Medicine, Southern Illinois University, USA*
Flavio Esposito[1] FLAVIO.ESPOSITO@SLU.EDU
[1] *Computer Science Department, Saint Louis University, USA*

Editors: Under Review for MIDL 2026

Abstract

Module diversity fundamentally enhances a model's ability to learn geometric structure by enabling a broader and more expressive set of feature representations. While many architectures improve performance by scaling parameters or relying on large-scale pretraining, these strategies make it difficult to identify which design principles truly enhance feature learning capability, especially in challenging domains with limited data such as laparoscopic surgical segmentation. This work investigates a parameter-constrained, no-pretraining setting to isolate the intrinsic feature learning capability of different module configurations. We introduce expert branches, a design concept that assigns different module families to their own independent pathways rather than mixing all features within a single stream. This separation encourages branch-specific specialization (Experts), reduces parameters, and avoids the entanglement that commonly obscures each module's contribution. We test this idea with TriEB, a UNet-based model incorporating CNN, deformable-convolution, and dynamic-snake branches with less total parameters. TriEB surpasses the vanilla UNet, the non-diverse TriCNN counterpart, and transformer-based models including SegFormer and Swin on the DSAD laparoscopic dataset. These results demonstrate that expert branches offer a more effective design principle for extracting diverse features from surgical imagery. The study highlights module diversity as a promising, architecture-agnostic framework for building efficient, interpretable, and data-adaptive feature extractors.

Keywords: Network design, Multi-Branch Network, Segmentation, Surgical Imaging.

1. Introduction

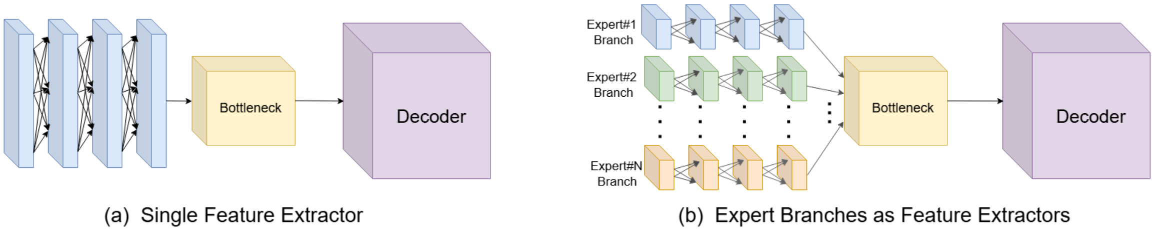
Feature extractors determine what geometric structure a vision model can perceive and how well it can generalize. This role becomes especially critical in laparoscopic image segmentation, where surgical scenes exhibit extreme visual complexity, including rapid camera motion, highly deformable organ surfaces, specular reflections, smoke, bleeding, tool occlusions, and thin, sharply curved anatomical structures. These conditions create severe appearance variability and weak global context, making segmentation fundamentally a problem of robust, geometry-aware feature learning under uncertainty. Yet no single feature extraction module is universally suited for such diverse geometric regimes: different architectural modules favor different structural patterns. Most existing architectures nevertheless collapse all features into a single dominant pathway. This motivates a feature extraction strategy that embraces module diversity by design, allowing heterogeneous extractors to specialize rather than compete.

To address this limitation, we propose Expert Branches (EB), a general architectural abstraction for building module-diverse feature extractors. An Expert Branch is defined as an independently operating feature-extraction pathway that maintains its own computational stream and exposes its output only at explicit fusion points. By preventing intermediate tensor sharing across branches, EB encourages each branch to cultivate unique and complementary representations rather than converging toward redundant filters—a common outcome when many channels originate from the same entangled layer. Importantly, EB imposes no restriction on the internal structure of each branch: any module type, network block, or architectural family may serve as an expert. This makes EB a general-purpose design principle rather than a convolution-specific construction.

In this work, our goal is not to chase the highest possible segmentation accuracy through architectural scale or external knowledge, but to isolate and verify the intrinsic effect of module diversity on feature learning. We therefore adopt a highly constrained experimental setup: (i) minimal structural change to the backbone, (ii) compact parameter budget, and (iii) no external pretraining. Under this controlled setting, any observed performance gain can be directly attributed to how features are learned, rather than to model size or transferred representations.

We conducted the verification experiment on a simple realization, TriEB: a UNet backbone for laparoscopic organ segmentation. TriEB assigns three different convolutional module types to three independent expert branches, which process the same input but evolve separately and interact only through lightweight fusion modules. This design ensures that feature diversity arises explicitly from module diversity, rather than from widened channels within a single transformation family. Importantly, the use of three convolution types here serves strictly as a methodological probe to validate the EB principle, not as a claim that these particular operators form an optimal set. Specifically, standard convolutions tend to capture stable textures and low-frequency patterns, deformable convolutions (DCN) adjust sampling positions to accommodate nonrigid or spatially misaligned anatomy (Dai et al., 2017; Zhu et al., 2019), and dynamic snake convolutions (DSC) incorporate geometric priors that emphasize curved or boundary-sensitive structures (Qi et al., 2023).

Transformer-based vision models have advanced rapidly since ViT showed that pure attention architectures can rival CNNs under large-scale pretraining (Dosovitskiy et al., 2021b). Subsequent studies in medical imaging consistently report that their strong performance depends heavily on extensive pretraining or large in-domain datasets (Shamshad et al., 2023; Takahashi and colleagues, 2024). In practice, however, surgical datasets are often small and highly domain specific, limiting the ability of ViT-style models to learn rich low-level geometric features from scratch. Although segmentation architectures such as SegFormer and Swin (Xie et al., 2021; Liu et al., 2021) achieve impressive results with strong pretraining, their global attention mechanisms do not inherently guarantee expressive boundary- or deformation-aware features under data-limited conditions, consistent with recent findings in medical self-supervised learning (Huang et al., 2023; Zeng and colleagues, 2024). Accordingly, all transformer baselines in this work are trained from scratch on DSAD (Carstens et al., 2023) to enable a fair comparison of intrinsic feature learning capacity. Pretrained transformer experts will be explored in future EB systems.



(a) Single Feature Extractor (b) Expert Branches as Feature Extractors

Figure 1: Structural comparison between (a) a conventional single-stream feature extractor that fully entangles all channels within one shared pathway, and (b) our Expert Branch architecture, which separates heterogeneous feature extractors into independent pathways.

When evaluated on the challenging DSAD laparoscopic dataset, TriEB consistently outperforms the vanilla UNet (Ronneberger et al., 2015), baseline multi-branch variants built from a single operator family, and transformer-based baselines trained from scratch, demonstrating a stronger ability to capture complex geometric cues directly from limited-domain data. The independent branches further exhibit interpretable specialization behavior, revealing how different experts contribute to distinct anatomical structures. Expert Branches provide a simple yet powerful abstraction for building module-diverse, data-adaptive feature extractors by explicitly separating heterogeneous feature learners into independent computational routes. TriEB serves as a minimal and controlled validation of this principle and lays the foundation for future EB systems that integrate broader architectural families—including pretrained transformer experts—for even stronger surgical scene understanding.

2. Related Work

Multi-branch architectures have been widely explored to enhance representational diversity, including Inception networks (Szegedy et al., 2015), frequency-decomposed hybrids, and convolution–attention mixtures such as ACmix (Pan et al., 2023). While these designs expand feature diversity, features from different operators are typically merged early, producing entangled representations that limit interpretability. Moreover, because branches interact through shared channel flows, pruning and scaling often require disruptive architectural changes.

In contrast, our work organizes heterogeneous operators into independent expert branches. By preserving the identity of each operator throughout the encoder rather than mixing features immediately, this design enables specialization, operator-aware analysis, and predictable, structurally safe pruning.

Laparoscopic and endoscopic images pose challenges distinct from those in conventional medical imaging. Real surgical scenes involve rapid camera motion, instrument occlusions, specular highlights, smoke, bleeding, and extreme nonrigid deformations. Anatomical structures often appear only partially and change appearance as instruments interact with tissue, while thin, tortuous boundaries such as vessels and ducts are particularly difficult to recover with a single operator family.

Most surgical segmentation pipelines are built upon convolutional networks, including UNet variants (Ronneberger et al., 2015; Çiçek et al., 2016). Standard convolutions provide efficient local feature extraction but are limited in modeling geometric variability. Deformable convolutions (DCN) (Dai et al., 2017; Zhu et al., 2019) improve robustness to shape changes through adaptive sampling, and Dynamic Snake Convolution (DSC) (Qi et al., 2023) embeds geometric priors to enhance curved and boundary-focused structures. These operators exhibit complementary strengths, yet existing systems usually adopt only one operator family or combine them within shared blocks where representations quickly become entangled, obscuring operator contributions and complicating structured pruning.

Vision transformers (ViTs) have emerged as a dominant paradigm in computer vision (Dosovitskiy et al., 2021a; Liu et al., 2021; Raghu et al., 2021; Li et al., 2024), but their strong performance relies heavily on large-scale pretraining. In medical and surgical imaging, where datasets are smaller and visually distinct, ViT models trained from scratch often underperform strong convolutional baselines, particularly on thin structures and irregular boundaries (Xie et al., 2021; Liu et al., 2021; Chen et al., 2023). This highlights a fundamental limitation: without extensive pretraining, transformers struggle to learn domain-specific geometric cues directly from limited data.

These limitations share a common root: effective surgical segmentation requires feature extractors that can jointly model stable regional appearance, large nonrigid deformation, and fine curvilinear boundaries. Yet most existing architectures either commit to a single dominant module type or entangle heterogeneous features too tightly to preserve their individual strengths. The Expert Branch strategy pursued in this work directly addresses

3. Methodology

3.1. Expert Branches: A Concept of Module Diversity

The core idea of this work is the principle of Expert Branches (EB), which organizes heterogeneous feature extraction modules into disentangled and independent pathways rather than mixing all feature channels within a unified shared stream. Different modules exhibit distinct inductive biases and are naturally suited to different geometric regimes. By separating them into dedicated branches, the model allows each module to express its representational strengths without interference from others. As a result, EB maintains a diverse and complementary set of geometric representations, rather than forcing all features to conform to a single transformation family.

The Expert Branch concept departs from conventional architectures that increase capacity by widening layers or deepening a single dominant module type. Such designs force all channels to propagate through the same feature transformation, even though many visual cues may not be well modeled by that module. Moreover, blending heterogeneous modules within the same block—such as through naive concatenation or tightly coupled attention-based fusion—entangles their effects, obscures their individual functional roles, and makes structural pruning unpredictable. Expert Branches instead establish clean, module-specific feature streams with explicit fusion points, enabling the network to learn how different experts should collaborate without artificially coupling their internal representations.

Parameter efficiency is another key motivation of the EB design. Because each branch is narrower than a full-width backbone encoder, the model avoids the quadratic parame-

Table 1: Approximate parameter breakdown (in millions). Totals match the overall model sizes: Vanilla UNet ≈ 31M, TriEB_CC ≈ 25M, TriEB_WF ≈ 19M.

Component	**Vanilla UNet**	**TriEB-CC**	**TriEB-WF**
Stem	0.11	0.09	0.09
Encoder Stage 1	0.74	0.66	0.42
Encoder Stage 2	2.95	2.12	1.58
Encoder Stage 3	11.80	8.45	5.80
Encoder Stage 4	11.80	8.45	5.80
Bottleneck	2.10	1.52	1.04
Decoder (all stages)	1.55	1.61	1.19
Fusion modules	0.00	1.20	0.20
Classifier head	0.09	0.09	0.09
Total	**31.0**	**24.9**	**18.8**

ter growth associated with uniformly widening convolutional layers. Even with multiple expert branches, the total parameter count remains comparable to or lower than that of the original single-stream backbone. At the same time, the diversity of modules creates a richer geometric basis for feature learning, allowing the network to capture structures that a single module family could not represent efficiently. This balance between reduced per-branch capacity and enhanced representational diversity underlies the practical strength of the Expert Branch framework.

3.2. TriEB: Minimal Expert Branch Instantiation, Bottleneck Behavior, and Fusion

We implement the Expert Branch concept within a UNet encoder to form TriEB, a minimal three-branch realization designed specifically for controlled methodological verification. In this setup, three different convolutional module types are assigned to three independent expert branches. This configuration is not intended to define an optimal operator combination, but rather to serve as a compact and interpretable experimental instantiation that introduces feature diversity with minimal structural change to the backbone. At each resolution level, all branches receive the same input feature map and process it independently without sharing intermediate activations. Interaction occurs only through lightweight fusion modules, preserving branch-level specialization while maintaining a compact encoder footprint.

In this initial study, we deliberately restrict TriEB to convolutional modules as a controlled testbed for validating the Expert Branch principle. Vision transformers are excluded as expert branches because their architectural form differs fundamentally from convolutions and would substantially increase the overall parameter budget, obscuring the contribution of module diversity itself. All models are trained from scratch to evaluate intrinsic feature learning capacity without external pretraining. Incorporating pretrained transformer modules as additional expert branches is a natural direction for future work.

Within this constrained setting, the three convolutional branches serve as one experimental realization for generating feature diversity through heterogeneous inductive biases. One branch uses standard convolutions (CNN) to model stable regional appearance, a second branch employs deformable convolutions (DCN) to adapt to nonrigid or spatially misaligned structures, and a third branch applies dynamic snake convolutions (DSC) to emphasize curved or boundary-sensitive patterns. To further isolate the effect of expert diversity, we conduct ablation studies using TriCNN, TriDCN, and TriDSC, where only a single operator type is used across all three branches under the same TriEB-WF fusion structure. The vanilla UNet baseline uses a base channel width of 64, while all TriEB-based models use a base width of 32.

Separating feature extractors into dedicated expert branches reshapes both the representational behavior and the capacity distribution of the encoder. Each branch is intentionally narrow: instead of allocating full-width channels to a single stream, TriEB redistributes capacity into multiple compact branches. Although each branch is narrower in isolation, their joint representation becomes more expressive. Before fusion, the concatenated feature tensor spans 96 channels, forming a broader intermediate bottleneck that preserves module diversity. This structural redistribution also leads to improved parameter efficiency. As shown in Table 1, TriEB reduces the parameter count from approximately 31M (vanilla UNet) to 25M in the channel-concatenation variant (TriEB-CC) and 19M in the weighted-fusion variant (TriEB-WF), while maintaining strong internal expressiveness. These properties also make Expert Branches naturally compatible with structured pruning and efficient scaling.

TriEB supports two fusion mechanisms at each encoder stage. The first, denoted TriEB-CC (channel concatenation), concatenates the three branch outputs,

$$\mathbf{Y}_{\mathrm{CC}}^{(s)} = [\,\mathbf{F}_{\mathrm{cnn}}^{(s)} \,\|\, \mathbf{F}_{\mathrm{dcn}}^{(s)} \,\|\, \mathbf{F}_{\mathrm{dsc}}^{(s)}\,],$$

followed by a 1 × 1 convolution for channel compression. This strategy preserves branch independence, maintains the largest bottleneck space, and is the most stable under pruning because feature streams remain disentangled. The second mechanism, TriEB-WF (weighted fusion), uses learned weights to combine the three branches,

$$\mathbf{Y}_{\mathrm{WF}}^{(s)} = \sum_{t\in\{c,d,s\}} g_t^{(s)}\,\mathbf{F}_t^{(s)}, \tag{1}$$

$$g^{(s)} = \mathrm{softmax}\left(W^\top \mathrm{GAP}\left([\mathbf{F}_t^{(s)}]\right)\right). \tag{2}$$

This fusion is more compact and parameter-efficient but introduces inter-branch dependencies, making structural pruning more sensitive.

Formally, at encoder stage s with input feature map $\mathbf{X}^{(s)}$, the three expert branches compute

$$\mathbf{F}_{\mathrm{cnn}}^{(s)} = f_{\mathrm{CNN}}^{(s)}(\mathbf{X}^{(s)}), \tag{3}$$

$$\mathbf{F}_{\mathrm{dcn}}^{(s)} = f_{\mathrm{DCN}}^{(s)}(\mathbf{X}^{(s)}), \tag{4}$$

$$\mathbf{F}_{\mathrm{dsc}}^{(s)} = f_{\mathrm{DSC}}^{(s)}(\mathbf{X}^{(s)}). \tag{5}$$

The fused output is then passed to the decoder, allowing complementary representations from heterogeneous experts to be combined adaptively for downstream segmentation.

3.3. Pruning as Expert Allocation

Rather than treating pruning as a post-hoc compression step, we integrate it into the EB framework as an *expert allocation* mechanism. All expert branches are initialized at full width and trained with mild sparsity regularization to encourage suppression of redundant channels. After convergence, channel saliency is estimated using the L1 norm of filter weights, followed by a single unstructured pruning pass that removes the lowest-scoring channels within each branch at a 30% pruning ratio. Because Expert Branches are architecturally independent and do not share intermediate activations, pruning does not introduce shape mismatches and can safely suppress weak channels or even entire branches. A brief fine-tuning stage of 25 epochs with a reduced learning rate is then applied to recover any accuracy loss. This process transforms pruning from a deployment optimization into a tool for analyzing the relative contributions of different experts.

3.4. Metrics

We report both region-based and boundary-based metrics to comprehensively evaluate segmentation performance.

Dice and mIoU (region-based). Dice and mean Intersection-over-Union (mIoU) quantify the pixel-wise overlap between predicted and ground truth masks and are widely used to assess volumetric segmentation accuracy. These metrics primarily reflect how well the model captures the overall extent of each anatomical region.

MASD (Mean Average Surface Distance). MASD measures the average geometric distance between the predicted and ground truth organ boundaries in both directions. It directly evaluates how far the predicted surface deviates from the true anatomical contour and provides a physically interpretable measure of boundary accuracy in pixel units (Maier-Hein et al., 2024).

NSD (Normalized Surface Dice). NSD measures the fraction of boundary points that lie within a predefined tolerance of the ground truth surface, normalized by the total boundary length (Maier-Hein et al., 2024). It reflects the proportion of the organ contour that is correctly localized within a clinically acceptable margin. In laparoscopic surgery, organ boundaries guide critical tasks such as dissection, exposure, and the preservation of delicate structures. While region-based metrics such as Dice and mIoU quantify volumetric correctness, they do not strongly penalize thin boundary deviations. In contrast, MASD and NSD directly evaluate how well a model captures fine anatomical contours—an essential requirement because organ boundaries are often thin, curved, and partially occluded; surgical instruments interact with tissues precisely along these boundaries; even small pixel-level errors can indicate incorrect delineation of vital structures; and many organs exhibit similar textures but distinct shape-driven signatures. Therefore, boundary-based metrics provide a more clinically meaningful assessment of segmentation quality in surgical scenes, where accurate delineation is crucial for downstream decision support and intraoperative guidance.

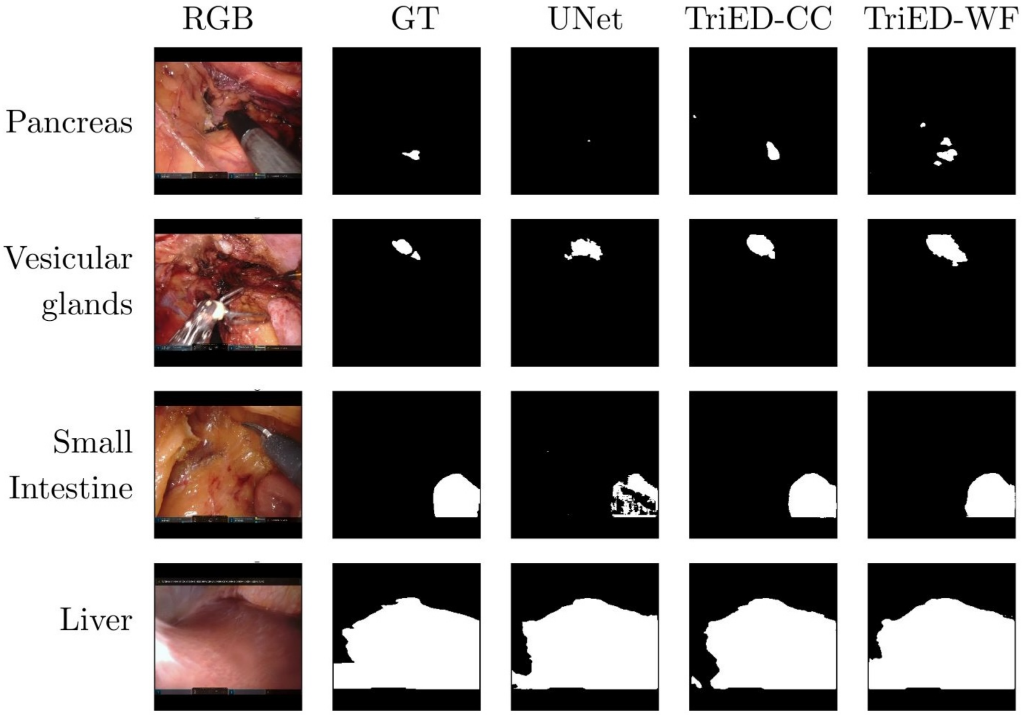


Figure 2: Qualitative comparison of the methods.The images are selected to highlight improvements.

4. Results

4.1. Overall Performance

Table 2 reports the average segmentation performance across all 11 DSAD organs, with qualitative examples shown in Fig. 2. Since all models are trained from scratch, the results directly reflect the intrinsic feature learning capability of each architecture without reliance on large-scale pretraining.

Both Expert Branch designs, TriEB-CC and TriEB-WF, outperform the vanilla UNet baseline and the transformer-based models (SegFormer and Swin) trained from scratch across nearly all metrics. The improvements are most consistent in mIoU, mDice, and NSD, demonstrating that explicit module diversity leads to stronger region-level accuracy and more reliable boundary localization. Compared to the single-operator ablation models (TriCNN, TriDCN, and TriDSC), the full TriEB designs achieve higher overall performance, confirming that the gains arise from heterogeneous expert collaboration rather than from any individual operator alone.

Boundary-focused metrics further highlight the advantage of Expert Branches. Both TriEB variants achieve the highest NSD among all models, indicating more precise anatomical contour alignment. This supports the motivation behind EB: independent experts capture complementary geometric cues that improve the representation of thin, curved, and deformable structures that dominate laparoscopic scenes. In contrast, transformer baselines trained from scratch show substantially weaker performance across both region-based

Table 2: Overall laparoscopic segmentation performance on DSAD averaged across 11 organs. Higher mIoU, mDice, and NSD and lower MASD indicate better performance. Ablation models (TriCNN, TriDCN, TriDSC) use a single convolution operator replicated across all branches to isolate the effect of module diversity. UNet serves as the single-stream baseline. Pruned models apply 30% unstructured pruning & fine-tune after convergence. Transformer baselines (SegFormer and Swin) are trained from scratch.

Model	**Setting**	**mIoU**	**mDice**	**MASD↓**	**NSD↑**
TriEB-WF	Pruned	0.456	0.542	27.133	0.449
TriEB-WF	Original	0.454	0.541	20.439	0.432
TriEB-CC	Pruned	0.471	0.559	25.442	0.467
TriEB-CC	Original	0.471	0.564	17.881	0.450
TriCNN (ablation)	Original	0.410	0.497	24.116	0.391
TriDCN (ablation)	Original	0.421	0.511	20.747	0.405
TriDSC (ablation)	Original	0.436	0.528	22.084	0.415
UNet (baseline)	Pruned	0.438	0.530	28.000	0.411
UNet (baseline)	Original	0.445	0.537	21.337	0.411
SegFormer (scratch)	Original	0.357	0.446	26.539	0.298
Swin (scratch)	Original	0.382	0.476	23.289	0.326

and boundary-based metrics, underscoring the difficulty of learning fine-grained geometric structure from limited surgical data without pretraining.

Pruning reveals additional insights into architectural robustness. For both TriEB variants, pruning has minimal impact on mIoU and mDice and slightly improves NSD, indicating that boundary localization remains stable under moderate sparsification. In contrast, the UNet baseline exhibits larger and less predictable changes after pruning, consistent with its fully entangled channel structure. Across models, MASD tends to increase after pruning, reflecting the emergence of occasional long-range false positives that strongly affect surface distance while minimally impacting region-based scores. This mixed behavior indicates that pruning sharpens true boundary representation while introducing isolated surface outliers.

4.2. Organ-Wise Evaluation

Table 3 reports organ-wise mIoU for all unpruned models trained from scratch. TriEB-CC achieves the highest accuracy in the majority of organs, with particularly strong improvements in anatomically complex structures such as the colon, stomach, liver, and small intestine. These organs exhibit steep curvature, variable shape, and strong nonrigid deformation—precisely the regimes where module diversity is expected to provide an advantage.

TriEB-WF performs competitively across most organs, indicating that even without maximizing the intermediate bottleneck space through full channel concatenation, allocating heterogeneous modules to independent branches remains effective. Both Expert Branch variants consistently outperform SegFormer and Swin when trained from scratch, reinforcing

Table 3: Organ-wise mIoU for unpruned models on DSAD. All models are trained from scratch.

Organ	UNet	TriEB-WF	TriEB-CC	SegFormer	Swin
Vesicular glands	0.217	0.215	0.216	0.114	0.122
Spleen	0.586	0.661	0.649	0.480	0.532
Colon	0.534	0.568	0.587	0.471	0.523
Small intestine	0.709	0.719	0.746	0.644	0.646
Stomach	0.484	0.542	0.561	0.392	0.415
Pancreas	0.187	0.192	0.210	0.141	0.162
Abdominal wall	0.758	0.760	0.769	0.712	0.708
Intestinal veins	0.361	0.315	0.328	0.279	0.283
Liver	0.516	0.603	0.608	0.430	0.492
Inf. mes. art.	0.266	0.269	0.296	0.188	0.211
Ureter	0.276	0.145	0.213	0.0810	0.105

the observation that transformer tokenization alone does not guarantee strong low-level geometric feature learning under limited data.

Several organ-level trends are evident. Expert Branches yield broad improvements across organs with diverse geometric characteristics, ranging from smooth surfaces (e.g., abdominal wall) to highly deformable and shape-complex regions (e.g., colon and stomach). TriEB models demonstrate stronger cross-organ robustness, reflecting improved geometric generalization from heterogeneous expert collaboration. While performance on extremely small and thin structures such as the ureter remains challenging for all methods, Expert Branches still provide competitive or improved performance relative to transformer baselines, highlighting the difficulty of boundary-sensitive segmentation under scratch training. All the evaluation confirms that Expert Branches improve intrinsic feature learning and geometric representation across a wide range of anatomical structures without relying on increased model scale or external pretraining.

5. Conclusion

Expert Branches provide a simple yet effective framework for promoting module diversity in feature extraction by allocating heterogeneous feature extractors to dedicated, disentangled pathways. This design moves beyond single-module backbones and avoids the representational entanglement that often obscures the functional role of individual components. We demonstrated this concept in TriEB using three convolutional expert branches as a minimal experimental realization within a UNet encoder. When trained from scratch on laparoscopic organ segmentation, TriEB consistently outperformed the vanilla UNet, single-operator ablation variants, and transformer-based models such as SegFormer and Swin, demonstrating the advantage of explicit module diversity under domain-limited conditions. The structural separation of branches also yielded stable behavior under moderate pruning. These results highlight Expert Branches as a promising direction for building efficient, interpretable, and data-adaptive feature extractors for challenging surgical scenes. Future work will explore larger datasets, pretrained expert modules, and extending the Expert Branch principle to more advanced surgical tasks.

Acknowledgments

This research is supported by...

References

Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):1–8, 2023.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. Springer, 2016.

Jieneng Chen et al. Vision transformers in medical imaging: A survey. *Medical Image Analysis*, 2023.

Jifeng Dai, Hao Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, and Thomas Unterthiner. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021a.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021b.

Shih-Cheng Huang, T. Kothari, et al. Self-supervised learning for medical image classification and segmentation: A review. *Insights into Imaging*, 14(1):55, 2023.

Yitong Li et al. Scaling vision transformers to 22b parameters. In *CVPR*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21 (2):195–212, 2024.

Xinglong Pan, Zhuang Jiang, Zehao Liu, Haoyi Zhang, Gao Huang, and Shiji Han. Acmix: A mixture of convolutional and attention branches for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11064–11074, 2023.

Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6070–6079, 2023.

Maithra Raghu et al. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Munawar Hayat Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802, 2023.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

Satoshi Takahashi and colleagues. Comparison of vision transformers and convolutional neural networks in medical image analysis: A literature review. *Journal of Medical Systems*, 48(4):105, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Xiaoyu Zeng and colleagues. Self-supervised learning frameworks for medical image analysis: A review. *Biomedical Engineering Online*, 23(1):12, 2024.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019.