Rethinking U-Net: Task-Adaptive Mixture of Skip Connections for Enhanced Medical Image Segmentation

Zichen Luo, Xinshan Zhu*, Lan Zhang, Biao Sun*

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, P. R. China
luozichen_01@tju.edu.cn, xszhu@tju.edu.cn, zl2022@tju.edu.cn, sunbiao@tju.edu.cn

Abstract

U-Net is a widely used model for medical image segmentation, renowned for its strong feature extraction capabilities and U-shaped design, which incorporates skip connections to preserve critical information. However, its decoders exhibit information-specific preferences for the supplementary content provided by skip connections, instead of adhering to a strict one-to-one correspondence, which limits its flexibility across diverse tasks. To address this limitation, we propose the Task-Adaptive Mixture of Skip Connections (TA-MoSC) module, inspired by the Mixture of Experts (MoE) framework. TA-MoSC innovatively reinterprets skip connections as a task allocation problem, employing a routing mechanism to adaptively select expert combinations at different decoding stages. By introducing MoE, our approach enhances the sparsity of the model, and lightweight convolutional experts are shared across all skip connection stages, with a Balanced Expert Utilization (BEU) strategy ensuring that all experts are effectively trained, maintaining training balance and preserving computational efficiency. Our approach introduces minimal additional parameters to the original U-Net but significantly enhances its performance and stability. Experiments on GlaS, MoNuSeg, Synapse, and ISIC16 datasets demonstrate state-of-the-art accuracy and better generalization across diverse tasks. Moreover, while this work focuses on medical image segmentation, the proposed method can be seamlessly extended to other segmentation tasks, offering a flexible and efficient solution for diverse applications.

Code — https://github.com/AshleyLuo001/UTANet

Introduction

Medical image segmentation is essential for analyzing and interpreting medical data, aiding healthcare professionals in diagnosing diseases, formulating treatment strategies, and tracking disease progression. Over the years, numerous segmentation models have been developed, such as convolutional Neural Networks (CNNs) are tailored for volumetric tasks, and Fully Convolutional Networks (FCNs) (Long, Shelhamer, and Darrell 2015) address specific challenges through multi-scale and skip connections. With U-Net (Ronneberger, Fischer, and Brox 2015) being one of the most widely adopted architectures due to its robust feature extraction capabilities and U-shaped structure. More and more variants(Peng, Sonka, and Chen 2023; Yu et al. 2023; Zhu et al. 2024) were developed. The skip connections in U-Net help recover spatial information lost during pooling operations. However, they fail to address the semantic gap between encoder and decoder, which becomes more pronounced with diverse datasets (Wang et al. 2022b). Different medical imaging tasks often require varying levels of semantic feature focus, which the simple skip connection mechanism in the original U-Net may fail to address adequately, leading to performance limitations. As a result, the direct transfer of features between the encoder and decoder without considering these differences can exacerbate the semantic gap, ultimately limiting the model's performance. Recent advancements in medical image segmentation have sought to overcome these limitations by refining the skip connection strategy to better align the semantic information between the encoder and decoder stages. For instance, UNet++(Zhou et al. 2018) employs nested and dense skip pathways to mitigate the semantic gap, but this comes at the cost of increased architectural complexity. Other approaches incorporate attention mechanisms(Oktay et al. 2018) or replace convolutional operations with transformers (Vaswani 2017; Chen et al. 2021; Zhang, Liu, and Hu 2021; Zheng et al.

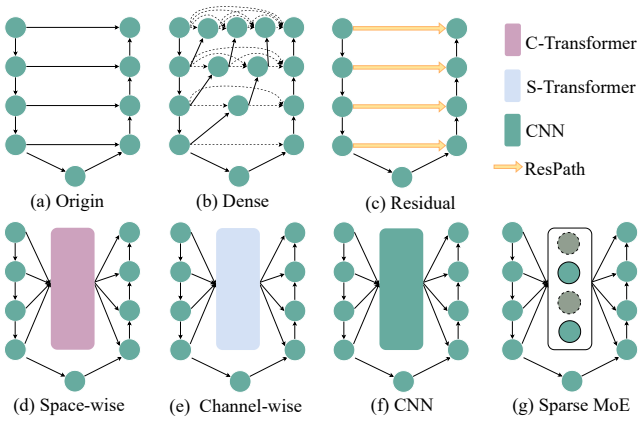

Figure 1: Different types of skip connections and the sparse skip connection we proposed.

*Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2021), which capture global context but often introduce substantial computational overhead. UCTransNet (Wang et al. 2022b) and UDTransNet (Wang et al. 2024) integrate attention mechanisms directly into skip connections to enhance feature weighting. While effective, these methods still adopt one-to-one connections and overlook the decoder's stage-specific preferences for semantic information, limiting their ability to generalize across datasets.

In this work, we explore different combinations of skip connection arrangements in U-Net through extensive experiments. As shown in Fig. 2, the optimal skip connection order consistently deviates from the original configuration. For example, on the GlaS dataset, U-Net achieves the highest accuracy when the skip connection order is 3f or 3421. We hypothesize that the information required by each decoder stage is not strictly one-to-one with the skip connections. Instead, the supplementary information for a particular decoding stage may need to originate from multiple skip connections. In other words, the decoder in U-Net exhibits specific information preferences for skip connections, which vary across different tasks or datasets. To tackle this issue, we propose a novel Task-Adaptive Mixture of Skip Connections (TA-MoSC) module that dynamically distributes the features passed through skip connections during the decoding phase. Inspired by the Mixture of Experts (MoE) framework (Jacobs et al. 1991), our approach introduces a task-adaptive mechanism that learns to combine encoder features from multiple stages to create optimal skip connections for each decoder stage. Unlike prior methods, our module considers the skip connections as a whole, leveraging the task allocation mechanism to respond to dataset-specific requirements of the decoder adaptively. We integrate the TA-MoSC module into the U-Net architecture, resulting in UTANet, which improves segmentation performance without significantly increasing model complexity. By employing a routing mechanism, UTANet generates task-specific skip connections, enabling the effective use of multi-scale features while minimizing information loss. This modular design is easily adaptable to various U-shaped architectures, making it suitable for diverse medical image segmentation tasks. Extensive experiments on public datasets demonstrate that UTANet achieves consistent improvements over baseline models, including absolute Dice score gains of 2.65% on GlaS, 6.55% on MoNuSeg, 3.82% on Synapse, and 1.35% on ISIC16. Additionally, we conduct an in-depth analysis of how feature interactions work. We summarize our main contributions as follows:

- We analyze skip connection configurations across diverse datasets and reveal that different tasks have different information preferences. Based on this, we redefine skip connections as a task allocation problem, using a dynamic scheduling mechanism to adaptively align multi-scale semantic features.
- We are the first to integrate the convolutional MoE framework into U-Net, introducing BEU to balance expert training, which achieves significant segmentation improvements with minimal parameters and high efficiency.
- Our method UTANet outperforms state-of-the-art methods on four datasets, with design principles generalizing to other domains and introducing a task-adaptive paradigm for segmentation.

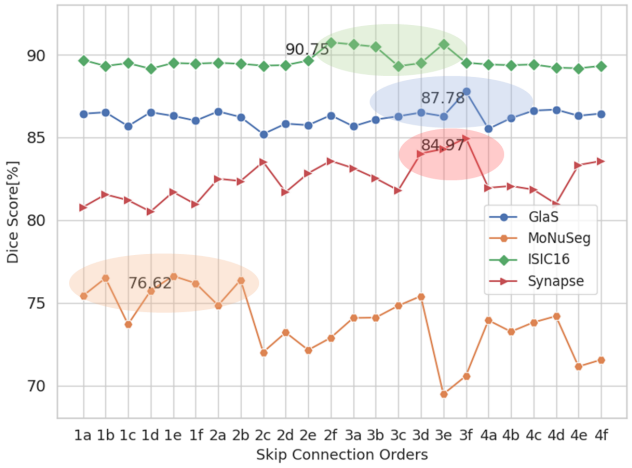

Figure 2: The effect of skip-connection using different combinations(permutations of 1234-4321) across different datasets.

Related Works

Advancements of U-Net Architectures

U-Net with its U-shaped structure and skip connections, remains foundational for segmentation tasks, facilitating the merging of high-resolution and context-rich features (Ronneberger, Fischer, and Brox 2015). Variants like ResUNet (Diakogiannis et al. 2020) and DenseUNet (Cai et al. 2020) introduced residual and dense connections to enhance feature reuse and flow. Attention mechanisms, as seen in Attention ResUNet (Li et al. 2021) and Attention U-Net (Oktay et al. 2018), improved focus on relevant features, while UNet++(Zhou et al. 2018) and UNet3+(Huang et al. 2020) tackled the semantic gap through nested skip pathways and multi-scale fusion, respectively. SelfReg-UNet (Zhu et al. 2024) improves the existing U-Net model by introducing two regularization mechanisms, SCR and IFD. Recently, Transformer-based architectures, such as TransUNet (Chen et al. 2021), Swin-UNet (Cao et al. 2022), and MCTrans (Ji et al. 2021), have leveraged the Vision Transformer (ViT)(Dosovitskiy et al. 2020) and its derivatives to enhance global context understanding. Techniques like gated axial attention (MedT(Valanarasu et al. 2021)) and cross-scale dependencies (MCTrans) further refined these models. However, despite their advancements, many analyses attribute the performance bottlenecks to the skip connections (Wang et al. 2022b). Methods such as UDTransNet (Wang et al. 2024) and EIU-Net (Yu et al. 2023) improved skip connections with attention-based recalibration, yet these approaches often lack the flexibility to adapt to diverse dataset requirements, emphasizing the need for task-adaptive mechanisms to align encoder and decoder features effectively.

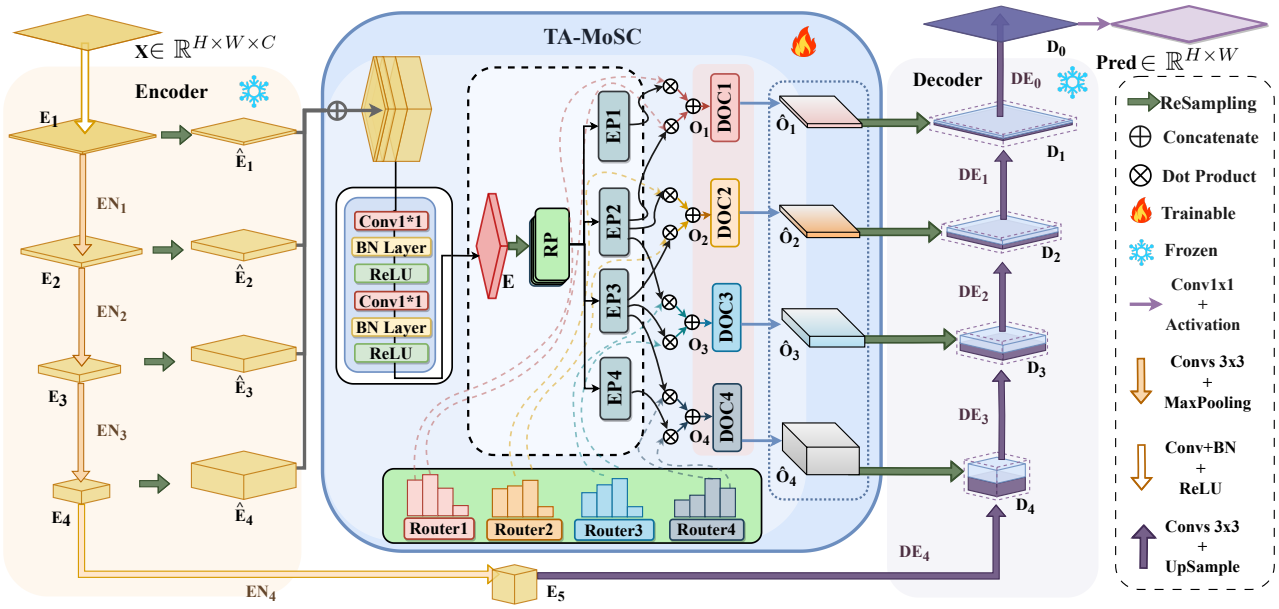

Figure 3: Illustration of the proposed UTANet. We replaced the original skip connections with our proposed TA-MoSC module, which uses a routing mechanism to deliver different feature information to each stage of the encoder. This approach helps bridge the semantic gap between the encoder and decoder.

Mixture of Experts (MoE)

Recently, sparse activation Mixture of Experts (MoE) (Jacobs et al. 1991) models has achieved remarkable success in scaling both visual (Cai, Gu, and Zhang 2018; Riquelme et al. 2021; Lou et al. 2021) and text models (Lepikhin et al. 2020a; Zoph et al. 2022). The primary motivation for using MoE is to increase model capacity without raising computational complexity (Shazeer et al. 2017) while controlling computational costs. Subsequently, the integration of MoE with transformer architectures (Lepikhin et al. 2020b; Fedus, Zoph, and Shazeer 2022) has further pushed the boundaries of network capacity. Additionally, MoE has proven effective in other challenging tasks. (Ma et al. 2018) addressed multi-task problems by designing a multi-gate MoE. Mustafa et al. (Zhai et al. 2022) utilized MoE to train a multi-modal expert mixture model based on contrastive learning. (Su et al. 2019) adopted each expert as an adapter, forming a task-specific adapter mixture to fine-tune a general image fusion framework. The sparsity of MoE reduces the risk of model overfitting and enhances multi-task learning performance by providing convenient inductive biases (Wang et al. 2022a). However, MoE has shown promise in tasks like natural language processing and vision but has yet to be explored in depth for image segmentation, particularly in the context of skip connections.

Method

Overview

In this paper, we propose UTANet, a UNet-based model incorporating Task-Adaptive Mixture of Skip-Connections (TA-MoSC), as illustrated in Fig. 3. This architecture combines dynamic expert selection, sparsity regularization, and a multi-gate mechanism for efficient computation, balanced resource use, and robust feature learning, making it ideal for multi-task learning and specialized feature processing. Given a medical image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$. After being encoded through multiple layers of the encoder, the multi-level encoder output features are fed into the TA-MoSC module for channel-wise feature distribution. This module outputs additional information features for the skip connections of each level decoder.

Task-Adaptive Mixture of Skip Connections

To address the issue that different task datasets have varying requirements for feature levels, our Task-Adaptive Mixture of Skip Connections(TA-MoSC) module leverages the superiority of MoE in task distribution to allocate the required semantic features to each level of the decoder. As shown in Fig. 4, the TA-MoSC consists of a router bank $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_4\}$, Skip-Connection(SC) bank, and four dockers.

Feature Aggregation Stripe To begin with, input image $\mathbf{X}$ is processed through multiple layers of an encoder, producing features at various levels, denoted as $\mathbf{E}_i \in \mathbb{R}^{\frac{H\times W\times C_i}{i^2}}, (i = 1, 2, 3, 4, 5)$:

$$EN = POOL_{max}(Conv(\cdot)), \tag{1}$$

$$\mathbf{E}_i = \begin{cases} ReLU(BN(Conv(\mathbf{X})), & \text{if } i = 1, \\ EN_{i-1}\left(\mathbf{E}_{i-1}\right), & \text{if } i > 1. \end{cases} \tag{2}$$

where, $POOL_{max}(\cdot)$ presents the maxpooling operation, where $Conv(\cdot)$ is the convolutional layer, $BN(\cdot)$ is the batch normalization function that standardizes the input of

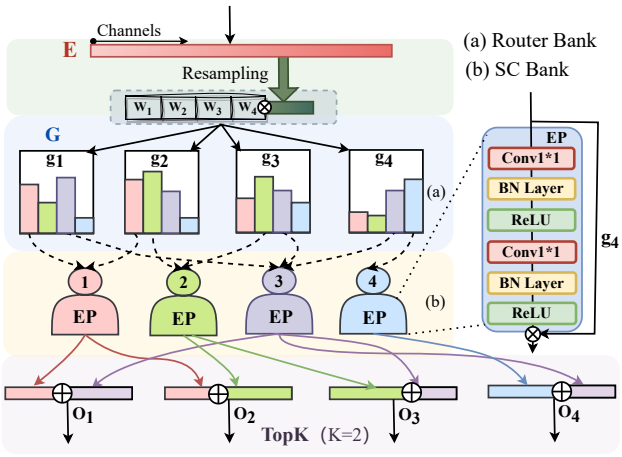

Figure 4: Routing phase process. A routing plan is generated for each decoder stage using shared expert models.

each mini-batch, mitigating issues like vanishing or exploding gradients, $ReLU(\cdot)$ is the ReLU activation function that introduces non-linearity. To ensure compatibility, we first resize all feature maps to the same size $(\frac{H}{2} \times \frac{W}{2})$ using bilinear interpolation $f_{bi}(\cdot)$. Only the first four feature levels $(i = 1, 2, 3, 4)$ are utilized for subsequent operations, as higher-level features $(i = 5)$ typically capture overly abstract representations that are less informative for detailed segmentation tasks, we scale the encoder features $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \mathbf{E}_4\}$ to the same size$\{\hat{\mathbf{E}}_1, \hat{\mathbf{E}}_2, \hat{\mathbf{E}}_3, \hat{\mathbf{E}}_4\}$. Then, we concatenate these resized features along the channel dimension to form a unified feature representation as feature aggregation stripe $\mathbf{E}$ that contains complete information and use $F_f(\cdot)$ defined as Eq. 4 to perform feature dimension reduction:

$$\begin{aligned} \hat{\mathbf{E}}_i &= f_{bi}(\mathbf{E}_i), i = 1, 2, 3, 4 \\ \mathbf{E} &= F_f(Cat(\hat{\mathbf{E}}_1, \hat{\mathbf{E}}_2, \hat{\mathbf{E}}_3, \hat{\mathbf{E}}_4)), \end{aligned} \tag{3}$$

$$F_f(\cdot) = ReLU(BN(Conv_{1\times1}(\cdot))), \tag{4}$$

where $Cat(\cdot)$ denotes the concatenation operation, $Conv_{1\times1}(\cdot)$ is the 1×1 convolution operation used for linear transformation of features while reducing the number of channels.

Routing Phase (RP) Now, we obtain the feature $\mathbf{E}$ aggregate multi-level semantic information. Next, the routing for skip connections at different stages will be specifically selected to customize the routing scheme $\mathbf{G}$ for the same input $\mathbf{E}$, each gate $\mathbf{g}_i$ generates a weight vector that assigns probabilities to all N experts:

$$\mathbf{g}_i = Softmax(POOL_{max}(\mathbf{E}) \cdot \mathbf{W}_i), \tag{5}$$

$$\mathbf{G} = [\mathbf{g}_i], i = 1, 2, 3, 4 \tag{6}$$

where $\mathbf{G}$ represents the selection probabilities for experts, and $\mathbf{W}_i$ is the learnable weight matrix for the gate. Next , we perform a weighted sum of the outputs from the experts to obtain the skip connections. Each router has a stage-specific preference for customizing the appropriate combination of experts and expert in our method is designed as an independent convolutional sub-network tasked with performing specialized nonlinear transformations on the input features allowing the model to adapt to diverse tasks or heterogeneous input distributions. Lightweight convolutional operations are used to maintain computational efficiency while preserving expressive power:

$$EP(\mathbf{E}) = \left[\text{ReLU} \circ \text{BN} \circ \text{Conv}_{1\times1}\right]^2(\mathbf{E}), \tag{7}$$

where, $\circ$ represents the composition of operations, the brackets denote the composition of functions, and the superscript "2" indicates that the composite operation is applied twice in succession.

After that, we obtain the router output $\mathbf{O}_i$ for each routing scheme:

$$\begin{aligned} \mathbf{q} &= [EP_n(\mathbf{E})]^T, n = 1, 2, \cdots, N \\ \mathbf{O}_i &= BN(\mathbf{G} \cdot TopK_i(\mathbf{q})), i = 1, 2, 3, 4 \end{aligned} \tag{8}$$

where $\mathbf{q}$ denotes all experts, and as shown in Fig.4, $TopK(\cdot)$ denotes top experts select operation while $TopK(\cdot)$ keeps only the top K(K=2) values, with the lowest probabilities set to 0. This sparse selection strategy ensures only a subset of experts is activated, reducing computational overhead.

Before decoders stage, we have obtained the results from various routing schemes. Next, we use the Docker module to shape and process each skip connection and then transport them to the corresponding decoders:

$$Doc(\cdot) = ReLU(Conv_{1\times1}(f_{bi}(\cdot))), \tag{9}$$

$$\hat{\mathbf{O}}_i = Doc(\mathbf{O}_i), \tag{10}$$

where $\hat{\mathbf{O}}_i$ is the final skip connection sent to the decoder, $Doc(\cdot)$ denotes the docker module.

Balanced Expert Utilization

The Balanced Expert Utilization Module is designed to enhance the efficiency and effectiveness of MoE models by ensuring that all experts are utilized during training. This module incorporates two key mechanisms: Expert Variance(EV) Loss and Unused Experts Handling.

Expert Variance Loss During training, the model may tend to over-rely on certain experts while neglecting others. This can lead to some experts being overworked while others are left underutilized, preventing each expert from fully contributing their capabilities. By calculating the variance in expert usage across different gating networks, we utilize Expert Variance(EV) Loss to encourage a more balanced utilization of experts:

$$\mathcal{L}_{EV} = \frac{1}{M} \sum_{p=1}^{M} \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i)^2}{\left(\frac{1}{N}\sum_{i=1}^{N} x_i + \epsilon\right)^2}, \tag{11}$$

$$\mathcal{L} = \beta \cdot \mathcal{L}_{EV} + \mathcal{L}_{DB}, \tag{12}$$

where M is the number of routers, N is the number of experts, and x_i represents each individual expert. ϵ is a small constant added to avoid division by zero and ensure numerical stability, set to 1×10^{-10}, β is a smoothing parameter, which we set to $\beta = 1 \times 10^{-3}$. L denotes the overall training loss, and $\mathcal{L}_{DB}$ represents the Dice-BCE Loss (Isensee et al. 2021), weighted by the Dice and the BCE losses.

Method	Param (M)	FLOPs (G)	MoNuSeg		GlaS		ISIC16	
			Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)
U-Net	14.8	50.3	76.45±1.86	62.87±2.01	85.45±1.26	74.48±1.66	90.28±0.41	83.72±0.46
UNet++	36.33	212.32	78.65±1.45	65.10±1.73	89.75±1.08	82.38±1.55	90.80±0.23	84.36±0.31
Att-UNet	34.88	105.28	77.95±1.95	64.72±2.13	90.65±0.31	83.69±0.51	90.74±0.24	84.45±0.26
Swin-Unet	41.38	17.38	76.69±0.96	62.43±1.23	88.27±1.63	75.38±1.68	90.63±0.17	84.04±0.15
TransUnet	105.28	49.34	76.91±0.68	62.70±0.85	87.42±0.98	79.16±1.31	90.71±0.26	84.29±0.26
R34UNet	24.37	47.82	78.94±1.73	65.56±1.62	91.14±0.32	84.41±0.48	90.43±0.15	84.01±0.22
Unet-v2	25.15	**8.27**	77.25±1.04	63.12±1.18	88.86±1.23	80.86±1.94	91.32±0.15	84.63±0.24
SelfReg-UNet	41.38	17.39	76.65±3.09	62.85±3.09	86.01±2.49	74.71±2.49	90.73±0.20	84.21+0.29
UDTransNet	33.90	53.02	78.88±1.73	65.24±1.78	90.16±0.72	82.08±1.14	91.04±0.21	84.78±0.30
Ours	**24.17**	64.41	**79.10±0.49***	**65.79±0.62**	**92.00±0.30***	**85.83±0.42**	**91.63±0.04**	**85.41±0.07**

Table 1: Quantitative results: The 5-fold cross-validation results on GlaS, MoNuSeg, and ISIC16 datasets to evaluate both the segmentation performance and stability of our model. The Dice and IOU are in 'mean ± std', boldface indicates the best-performing results, and * represents $p < 0.05$.

Unused Experts Handling This mechanism addresses the problem of experts being completely ignored during training. When the gating network does not select certain experts, a random sample from the input data is processed by these unused experts. This approach prevents experts from being idle and ensures that every expert is trained, promoting more balanced utilization and enhancing overall model performance, we first confirm the unused experts:

$$P_{un} = \{0, 1, ..., N-1\} \setminus \{top_k(\mathbf{g}_i, k)\}, \tag{13}$$

where N is the number of experts, P_{un} is the index set of unused experts and $top_k(\cdot)$ denotes the index of selected the first k experts. Then, we use the random sample to calculate the output of unused experts and update weightings:

$$y = \sum_i \mathbf{g}_i \cdot \mathbf{q}_i(\mathbf{x_r}) + \sum_j \mathbf{g}_i \cdot \mathbf{q}_j(\mathbf{x_r}), \tag{14}$$

where i is the number of used experts, j is the number of unused experts, and $\mathbf{x_r}$ is one random sample from the training dataset, $\mathbf{y}$ is the combined output from all experts. Finally, the redistributed skip connection information is supplemented to the decoder, which progressively decodes it to produce the final segmentation result:

$$\mathbf{D}_i = \begin{cases} DE_i(\mathbf{E}_{i+1}), i = 4 \\ DE_i(Cat(\hat{\mathbf{O}}_{i+1}, \mathbf{D}_{i+1})), i = 0, 1, 2, 3 \end{cases} \tag{15}$$

$$\mathbf{Pred} = Seg(\mathbf{D_0}), \tag{16}$$

where $\mathbf{Pred}$ denotes the final prediction after UTANet as shown in Fig.3, $DE_i(\cdot)$ represents the different decoders, and $\mathbf{D_i}$ refers to the features at each decoder stage. $Seg(\cdot)$ denotes the segmentation head.

Experiments

Datasets

We evaluate our model's performance and generalization ability using four medical image datasets, encompassing both small and large-scale scenarios. For smaller datasets, GlaS (Sirinukunwattana et al. 2017) contains 165 high-resolution H&E-stained images (85 for training and 80 for testing), while MoNuSeg (Kumar et al. 2017) comprises 44 images (30 for training and 14 for testing). For larger datasets, Synapse (Landman et al. 2015) includes 30 abdominal CT scans with 3,779 axial images covering 8 organs, and ISIC16 (Gutman et al. 2016) features 1,279 dermoscopic images (379 in the test set) with ground truth annotations for two disease categories. These datasets provide a thorough evaluation of our model across varying data scales and complexities.

Method	Dice ↑	HD95 ↓
U-Net	84.41±1.72	48.05±2.74
R34UNet	87.26±0.74	33.94±1.22
UNet++	85.65±0.92	37.45±0.29
Sin-UNet	86.78±1.65	31.66±1.58
Att-UNet	86.62±1.69	37.85±1.31
TransUNet	86.45±0.93	31.28±0.87
UDTransnet	86.93±1.17	37.83±0.31
Ours	**88.23±0.87**	**28.13±0.21**

Table 2: Quantitative results. The 5-fold cross-validation results on Synapse dataset. The Dice and 95% Hausdorff are in 'mean ± std', boldface indicates the best-performing results.

Implementation Details

Experiment Settings We train our model using PyTorch on a single NVIDIA 3090 GPU with 24GB of memory. To prevent overfitting, we applied online data augmentation techniques during training, such as horizontal and vertical flips, along with random rotations. After thorough testing, we determined that a batch size of 4 following (Valanarasu et al. 2021) was optimal for both the GlaS and MoNuSeg datasets. Similarly, for the Synapse dataset, we set the batch size to 8, and for the ISIC16 dataset, we used a batch size of 16. The input resolution for all datasets was standard-

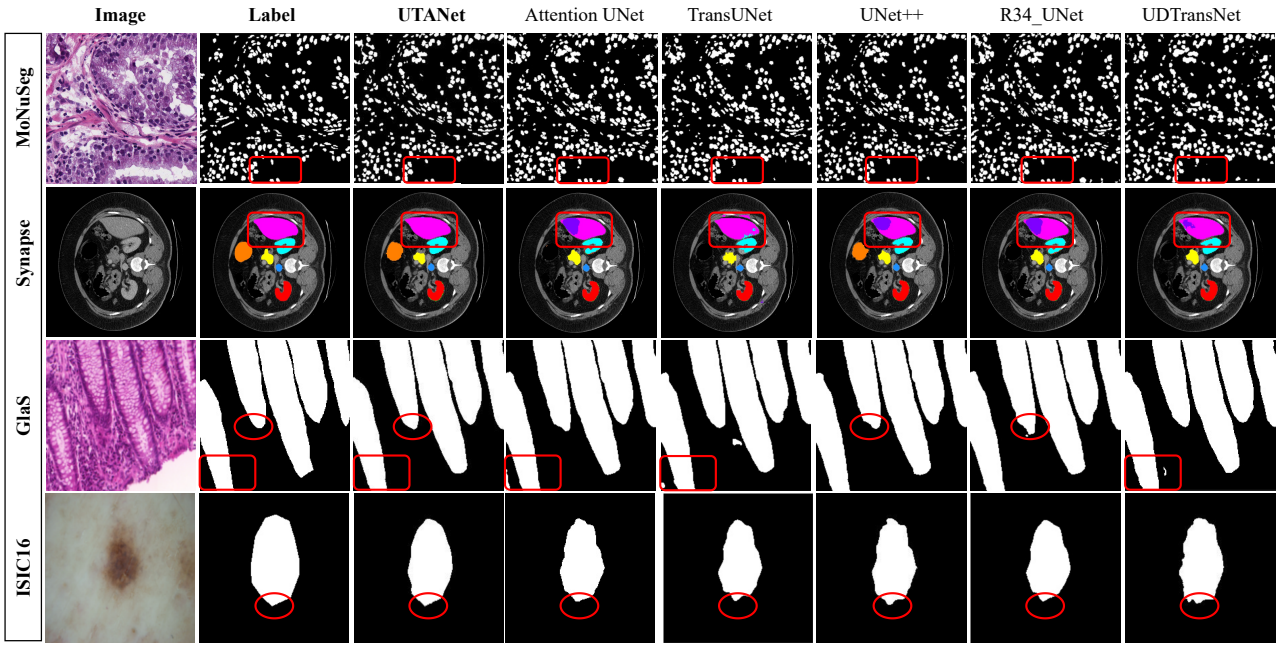

Figure 5: In the qualitative analysis across different datasets, the areas highlighted with red boxes indicate regions where our UTANet model performs better in segmentation compared to other models. It is evident that our proposed model achieves superior segmentation results.

Dataset	TAM	EVL	UEH	Dice(%)	IoU(%)
Glas	Baseline			85.45±1.26	74.48±1.66
	✓			89.77±2.59	82.40±3.56
	✓	✓		91.91±0.35	85.62±0.56
	✓		✓	90.96±0.83	84.37±1.13
	✓	✓	✓	**92.00±0.30**	**85.83±0.42**
MoNuSeg	Baseline			76.45±1.86	62.87±2.01
	✓			77.81±0.98	64.02±1.30
	✓	✓		78.61±1.05	65.16±1.28
	✓		✓	78.78±1.01	65.43±1.24
	✓	✓	✓	**79.10±0.49**	**65.79±0.62**

Table 3: Ablation studies on UTANet. TAM presents the TA-MoSC module which we proposed while EVL and UEH present expert variance loss and unused experts handling. The baseline model is U-Net.

ized to 224×224 pixels. The training was conducted with the Adam optimizer, starting with an initial learning rate of 0.001. The number of iterations was not fixed, thanks to the implementation of an early stopping strategy. We adjusted the learning rate over time using a Cosine Annealing schedule (Loshchilov and Hutter 2016). To further improve network performance, we employed a hybrid loss function that combines Cross-Entropy loss, Dice loss, and our custom-designed EV loss. It is important to note that all baseline models were trained using the same settings.

Training Strategy Our training process is divided into two stages. In the first stage, we use the original skip connections to train both the encoder and decoder, allowing them to become familiar with the dataset and acquire the basic encoding and decoding capabilities required for it. Once this is achieved, we freeze the encoder and decoder and focus on training our proposed TC-MoSC module. This enables each expert to fully learn the features they are specialized in, while the router, with a better understanding of the semantic requirements at each decoding stage, can effectively allocate tasks. At different decoding stages, different combinations of experts are weighted to provide the supplementary information needed for that specific stage. The primary purpose of freezing the encoder and decoder is to prevent their parameters from changing during the full training process, which would decrease the difficulty of training and make convergence easier to achieve.

Evaluation

In this section, we compared nine methods for improving U-Net, which can be categorized into two types: improvements in the encoder-decoder structure, including R34-UNet, R2UNet, UNet-v2, and SwinUNet, and improvements in the skip connections, including Attention U-Net, UNet++, SelfReg-UNet, and UDTransNet.

Quantitative Comparisons. We quantitatively assessed the effectiveness of our model using the Dice coefficient (Dice) and Intersection over Union (IoU) as performance metrics on the GlaS, MoNuSeg, and ISIC16 datasets. For the multi-label segmentation Synapse dataset, we used the 95% Hausdorff Distance metric. We evaluated model performance using 5-fold cross-validation (Kohavi 1995). Statistical significance was assessed with independent Student's t-tests (Student 1908) at $\alpha = 0.05$ (Fisher 1925). Our method

TopK	MoNuSeg	GlaS	ISIC16	Synapse
	Dice(%)	Dice(%)	Dice(%)	Dice(%)
1	77.87±0.83	89.05±0.77	78.89±6.64	45.41±7.49
2	78.97±0.61	91.85±0.13	90.41±0.73	82.13±7.60
3	**79.10±0.49**	**92.00±0.20**	**91.87±0.07**	86.83±**0.65**
4	78.90±0.63	91.18±0.65	91.63±0.14	**88.09**±0.92

Table 4: The settings for the TopK value in the inference stage. The best setting is 3 for single-label segmentation tasks, while for multi-label segmentation, 4 is the optimal setting.

significantly outperformed baselines on GlaS (p=0.003, Cohen's d=1.2) and MoNuSeg (p=0.002, Cohen's d=1.5), with large effect sizes (Cohen 1988). The Benjamini-Hochberg correction (Benjamini and Hochberg 1995) confirmed the robustness of these findings. As shown in Tab.1, the small standard deviations observed in our experimental results indicate that our model exhibits high stability and robustness, consistently performing well across different data splits and experimental conditions. Our method exceeds the performance of general U-Net variants, demonstrating superior compatibility across different datasets.

Qualitative Comparisons. The qualitative results for all four datasets are shown in Fig. 5. The red boxes highlight areas where UTANet outperforms other models. It is evident that UTANet captures detailed information more effectively on the MoNuseg dataset, which benefits from the detail-sensitive expert modules in our proposed TA-MoSC module. These four expert modules extract features at different levels and provide supplementary information to each stage of the encoder through task routing.

Ablation Studies

Ablation Studies on Proposed Modules As shown in Tab. 3, Baseline+TA-MoSC+BEU consistently outperforms other baselines. This indicates that incorporating our TA-MoSC module significantly improves the model's segmentation performance. Additionally, applying our proposed BEU optimization further enhances performance. Our results demonstrate that different datasets require varying levels of supplementary information during the decoding stage. By supplying the decoder with the appropriate supplementary information, the model's segmentation performance can be substantially improved.

Ablation Studies on TopK Settings As shown in Tab. 4, in the inference phase, we experimented with different TopK settings and found that setting TopK to 3 yielded the best test results on single-label datasets, while setting TopK to 4 on the multi-label Synapse dataset. Because on single-label datasets, when TopK is set to 1 or 2, the lack of supplementary information during the decoding stage leads to a decline in model performance. Conversely, when TopK is set to 4, it introduces unfavorable information that needs to be filtered out, also resulting in decreased performance. Therefore, setting TopK to 3 enables a weighted summation of outputs from three experts, capturing more comprehensive and beneficial information while filtering out harmful information.

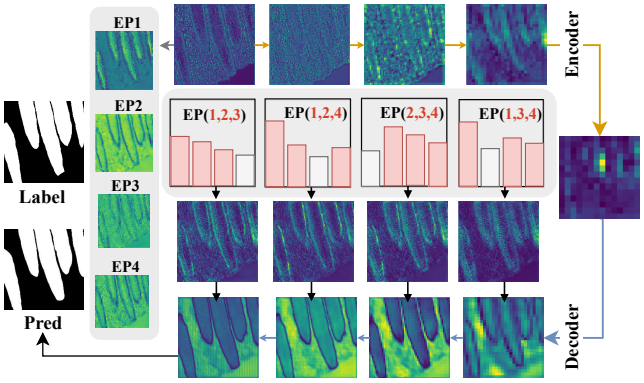

Figure 6: Visualize the operation mechanism of the expert module on GlaS dataset. The red text part is the prioritization of three different experts at every stage.

However, the multi-label dataset needs more channel information to classify the different multi-channels. For multi-label datasets, setting TopK to 4 allows for a weighted summation of outputs from all experts to obtain the full information needed for classification. We visualized the experts in the model trained on GlaS dataset. As shown in Fig. 6, the features and semantics learned by each expert network differ, after passing through the TA-MoSC module, the highlighted regions of the skip connection features align closely with those of the encoder's third and fourth stages. This finding resonates with our previous analysis of the UNet various orders which revealed that the final output stage exhibits a preference for semantic information predominantly derived from the encoder's third and fourth stages. This corroborates our initial finding that for different datasets or tasks, the model exhibits specific preferences for the information carried by skip connections. Therefore, our proposed TC-MoSC module effectively addresses the initial issue of information asymmetry in skip connections by precisely leveraging different expert networks to learn various semantic information.

Conclusion

During the decoding process of U-Net, the need for supplementary information varies across different datasets. This work bridges the gap by integrating MoE into skip connections, enabling adaptive redistribution of multi-scale features tailored to the decoder's stage-specific preferences. Unlike traditional fixed skip connection mechanisms, the proposed Task-Adaptive Mixture of Skip Connections (TA-MoSC) module dynamically aligns encoder and decoder features, addressing dataset-specific segmentation challenges. While this work focuses on medical image segmentation as a case study, the proposed approach can be generalized to other dense prediction tasks, such as natural image segmentation and object detection. In future work, we will focus on optimizing the computational efficiency of the model, which will allow the TA-MoSC module to be more scalable and efficient, ensuring its practicality for real-time applications and large-scale tasks across various domains.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grants No. T2322020, 62371329 and 61972282.

References

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1): 289–300.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062.

Cai, S.; Tian, Y.; Lui, H.; Zeng, H.; Wu, Y.; and Chen, G. 2020. Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quantitative imaging in medicine and surgery*, 10(6): 1275.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Diakogiannis, F. I.; Waldner, F.; Caccetta, P.; and Wu, C. 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 94–114.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd.

Gutman, D.; Codella, N. C.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; and Halpern, A. 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1605.01397*.

Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; and Wu, J. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 1055–1059. IEEE.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Ji, Y.; Zhang, R.; Wang, H.; Li, Z.; Wu, L.; Zhang, S.; and Luo, P. 2021. Multi-compound transformer for accurate biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, 326–336. Springer.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1137–1145. Morgan Kaufmann.

Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; and Sethi, A. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7): 1550–1560.

Landman, B.; Xu, Z.; Igelsias, J. E.; Styner, M.; Langerak, T.; and Klein, A. 2015. Segmentation outside the cranial vault challenge. In *MICCAI: multi Atlas labeling beyond cranial vault-workshop challenge*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020a. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020b. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, R.; Zheng, S.; Duan, C.; Su, J.; and Zhang, C. 2021. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lou, Y.; Xue, F.; Zheng, Z.; and You, Y. 2021. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1930–1939.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Peng, Y.; Sonka, M.; and Chen, D. Z. 2023. U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation. *arXiv preprint arXiv:2311.17791*.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR*, abs/1505.04597.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Sirinukunwattana, K.; Pluim, J. P.; Chen, H.; Qi, X.; Heng, P.-A.; Guo, Y. B.; Wang, L. Y.; Matuszewski, B. J.; Bruni, E.; Sanchez, U.; et al. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35: 489–502.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1): 1–25.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Valanarasu, J. M. J.; Oza, P.; Hacihaliloglu, I.; and Patel, V. M. 2021. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, 36–46. Springer.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wang, D.; Liu, J.; Fan, X.; and Liu, R. 2022a. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022b. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2441–2449.

Wang, H.; Cao, P.; Yang, J.; and Zaiane, O. 2024. Narrowing the semantic gaps in u-net with learnable skip connections: The case of medical image segmentation. *Neural Networks*, 106546.

Yu, Z.; Yu, L.; Zheng, W.; and Wang, S. 2023. EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation. *Computers in Biology and Medicine*, 162: 107081.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18123–18133.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, 14–24. Springer.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11. Springer.

Zhu, W.; Chen, X.; Qiu, P.; Farazi, M.; Sotiras, A.; Razi, A.; and Wang, Y. 2024. SelfReg-UNet: Self-Regularized UNet for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 601–611. Springer.

Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; and Fedus, W. 2022. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*, 2(3): 17.