

## Article

# HT-Net: A Task-Specific Dual-Branch Network for Tumor Segmentation in Histopathological Images

Lijuan Ren <sup>1</sup>, Hang Chen <sup>1,\*</sup>, Jun Yang <sup>2</sup>, Lei He <sup>1,\*</sup> and Dan Tang <sup>1</sup><sup>1</sup> School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China<sup>2</sup> Sichuan Provincial Big Data Technology Services Center, Chengdu 610020, China

\* Correspondence: 3240707065@stu.cuit.edu.cn (H.C.); helei1978@cuit.edu.cn (L.H.)

## Abstract

Breast cancer, the most prevalent malignancy among women globally, requires precise assessment of lymph node metastases in whole-slide images (WSIs) for accurate clinical staging and therapeutic planning. However, manual assessment is laborious and prone to inter-observer variability. Inspired by biomimetic principles of functional specialization and complementary visual perception, we propose Head–Tail Network (HT-Net), a novel dual-branch encoder network for histopathological image segmentation. HT-Net employs a HeadBranch with Squeeze-and-Excitation (SE) attention to prioritize background feature selection and a TailBranch with a novel Hybrid Multi-Scale Attention (HMSA) module to resolve complex tumor boundaries. Features from both branches are integrated through a spatially adaptive fusion module and reconstructed using a U-Net-style decoder. A task-specific supervision loss further enforces functional specialization between branches and mitigates class imbalance. Experiments demonstrate that HT-Net outperforms U-Net, UNet++, UCTransNet, VM-UNet, and nnWNet. On Camelyon16, HT-Net achieved 89.37% mDice and 82.05% mIoU, surpassing nnWNet by 0.86 and 1.08 percentage points, respectively. On Glas, it achieved 93.04% mDice and 86.95% mIoU, exceeding nnWNet by 0.30 and 0.47 percentage points. These results demonstrate the effectiveness of the biomimetic-inspired functional specialization strategy for automated histopathological image segmentation.

**Keywords:** tumor segmentation; histopathological imaging; dual-branch encoder; multi-scale attention; task-specific supervision



Academic Editors: Ameer Tamoor Khan, Shuai (Steven) Li and Bolin Liao

Received: 23 August 2026

Revised: 9 October 2026

Accepted: 12 October 2026

Published: 14 October 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

Breast cancer remains the most prevalent malignant neoplasm among women worldwide. Global cancer statistics indicate that in 2020, there were approximately 2.26 million new cases, surpassing lung cancer as the leading cause of cancer-related mortality among females [1,2]. Because lymph node metastasis significantly dictates clinical staging and therapeutic strategy, the precise evaluation of sentinel lymph nodes (SLNs) has become a cornerstone of pathological diagnosis [3]. Currently, pathological analysis utilizing whole-slide imaging (WSI) is the gold standard for assessing SLN status [4]. However, this approach relies heavily on the manual, region-specific examination by specialized pathologists. A single WSI can contain hundreds of thousands of pixels, and a comprehensive inspection is labor-intensive and time-consuming. Furthermore, diagnostic consistency is often compromised by a pathologist's subjective experience and fatigue, leading to a non-negligible risk of misdiagnosis or overlooked metastases [5]. Meanwhile, these efficiency bottlenecks are

exacerbated by a global shortage of pathologists. Moreover, the acquisition of high-quality, expert-annotated datasets is prohibitively expensive [6], creating a data bottleneck for deep learning applications. Consequently, automated pathological image analysis enabled by deep learning has gradually emerged as a research frontier [7], offering a promising pathway to enhance diagnostic efficiency and mitigate subjective bias. These challenges highlight the need for automated analysis methods capable of efficiently processing complex pathological visual information while preserving clinically meaningful morphological characteristics. In this context, biomimetics provides a useful conceptual framework for developing intelligent visual analysis systems. Biological perception commonly relies on hierarchical processing, selective attention, and functional specialization to extract complementary information from complex environments. Translating these principles into computational architectures offers a promising strategy for handling the heterogeneous and highly structured visual information encountered in histopathological images.

Early investigations into automated analysis primarily adopted the patch-level classification paradigm, where WSIs are segmented into fixed-size local regions for individual assessment. For instance, Wang D et al. [3] employed this approach to achieve pixel-level detection of lymph node metastasis. Subsequently, end-to-end methodologies utilizing Multiple Instance Learning (MIL) sought to capture global context; Wang Z et al. [8] introduced a Transformer-guided MIL framework that characterizes contextual relationships among patches via a self-attention mechanism and leverages transfer learning to its predictive accuracy, even with scarce labeled data. Nonetheless, the paradigms—whether patch-based or MIL-based—fundamentally generate “region-level” or “slice-level” probabilities. Such outputs fail to precisely represent tumor morphological contours and invasion boundaries [9]. From a biomimetic perspective, this limitation can be viewed as a consequence of relying primarily on a single level of visual representation, whereas biological perception achieves robust recognition through complementary processing of global context and fine-grained local information. This technique in SLN diagnosis, clinicians require more than a binary “presence or absence” result, they must evaluate tumor morphology, distribution, and extent—features that are difficult to capture with weakly supervised MIL or patch-level classification. Therefore, moving toward pixel-level segmentation not only provides finer pathological characterization but also creates an opportunity to incorporate biomimetic principles of complementary feature extraction and adaptive information integration into network design. As a result, research has shifted toward the more complex task of tumor segmentation.

Tumor segmentation in histopathological images aims to assign a categorical label (e.g., tumor, stroma, lymphocytes) to each pixel, providing clinicians with a comprehensive “tumor distribution heatmap.” This technique enhances analytical granularity from the “block level” to the “pixel level”, enabling a more precise delineation of tumor extent. The encoder–decoder architecture has become the predominant framework for this task, owing to its comprehensive pixel prediction functionality. Specifically, U-Net [10]—with its symmetric architecture and skip connections—effectively addresses the challenge of limited training data in medical imaging. Subsequent advancements have expanded this framework: UNet++ introduced nested dense skip connections to bridge the semantic gap [11], Attention U-Net utilized gated attention to highlight fine-scale features [12], and TransUNet [13] and Swin-UNet [14] integrated Transformers to enhance global modeling. While attention modules like SE-Net [15] and CBAM [16] have further improved feature discriminability in both channel and spatial dimensions, a core contradiction remains unresolved in current methodologies.

Existing methods typically rely on the implicit assumption that features across different image regions can be uniformly extracted through a single encoding strategy. In

pathological tumor segmentation, this assumption is often invalid due to three primary challenges: (1) tumor regions and adjacent normal tissue exhibit vast differences in texture, color, and morphology, which a single-branch encoder struggles to capture simultaneously. (2) tumor regions often occupy only a small fraction of the slide area, causing the supervisory signal to be dominated by the background and constraining the model's perception of tumorous areas. (3) the transition between tumors and normal tissue is often characterized by gradual and blurred edges, which are difficult for static receptive fields to resolve.

To address these challenges, we introduce Head–Tail Network (HT-Net), a dual-branch encoder network that optimizes tumor segmentation through innovations in architectural design, feature modeling, and supervision. The primary contributions of this study are as follows:

1. **Dual-Branch Parallel Encoder:** We propose an architecture comprising a HeadBranch (background extraction) and a TailBranch (tumor extraction). These branches utilize distinct attention mechanisms tailored to their respective regions, replacing conventional homogeneous extraction to address regional feature heterogeneity.
2. **Hybrid Multi-Scale Attention (HMSA):** We developed an HMSA module that integrates parallel channel attention pathways with varying compression ratios and spatial attention branches using dilated convolutions. This allows for the adaptive representation of channel dependencies and spatial context across multiple scales, significantly enhancing boundary detection.
3. **Task-Specific Supervision:** We implement a specialized loss function where the HeadBranch receives auxiliary supervision focused on background regions, while the TailBranch employs multi-scale deep supervision via spatial weight maps. This approach enforces functional specialization between branches and mitigates class imbalance.
4. **State-of-the-Art Performance:** Comprehensive evaluations on the Camelyon16 and Glas benchmark datasets demonstrate that HT-Net outperforms current state-of-the-art models, confirming its effectiveness for histopathological image segmentation.

### 1.1. Related Work

#### 1.1.1. Medical Image Segmentation Methods

The rapid evolution of deep learning has catalyzed transformative advancements in medical image segmentation. U-Net [10] remains a seminal contribution to this field, establishing the groundwork for modern medical image segmentation techniques through the introduction of a symmetric encoder–decoder architecture and a skip connection mechanism. Subsequent research has expanded upon this baseline from various perspectives to address specific diagnostic challenges. Zhou et al. [11] proposed UNet++, which outperforms the traditional U-Net in multiple lesion segmentation tasks by reducing the semantic discrepancy between encoders and decoders through layered dense skip connections. Alom et al. [17] developed R2U-Net, integrating residual connections and recurrent convolutions to augment feature extraction efficacy for retinal vascular and skin lesion segmentation. To facilitate multi-scale feature fusion, Huang et al. [18] introduced UNet3+, utilizing full-scale skip connections and deep supervision techniques to improve contour delineation in organ CT segmentation. Furthermore, Ibtehaz et al. [19] proposed MultiResUNet, which reconstructs convolutional pathways using multi-resolution blocks and residual paths to bridge the semantic gap across adjacent scales, demonstrating robust generalization in multimodal biomedical tasks.

In recent years, Transformers have exhibited significant efficacy in medical image segmentation due to their global self-attention mechanism. Chen et al. [13] introduced TransUNet, which integrates a Vision Transformer into the bottleneck layer of a U-shaped architecture, substantially surpassing convolutional CNN-based approaches in abdominal

multi-organ CT segmentation. Similarly, Cao et al. [14] proposed Swin-UNet, a U-shaped architecture composed entirely of Transformers, which attained superior accuracy on the Synapse dataset, particularly for small-scale organs like the pancreas. To address regions with indistinct boundaries, Lin et al. [20] presented DS-TransUNet, which incorporates a dual-scale Swin Transformer encoder to enhance segmentation precision.

Despite these significant advancements, existing methodologies typically rely on a single-branch encoder architecture. These models apply a uniform feature extraction strategy to both Tumor regions and normal background areas, which hinders the model's ability to learn distinct, region-specific representations. This limitation serves as the fundamental motivation for the dual-branch encoder architecture proposed in this study.

### 1.1.2. Medical Image Segmentation Based on Attention Mechanisms

Attention mechanisms have emerged as a fundamental technology for improving the feature representation capabilities of medical image segmentation networks. Hu et al. [15] introduced SE-Net, which models inter-channel dependencies via “squeeze-and-excitation” operations that adaptively recalibrate channel-wise feature responses. Although this method efficiently suppresses extraneous textures and enriches the semantic representation of lesions—particularly in dermoscopic image segmentation. However, it exclusively emphasizes global information at the channel level and fails to recognize the spatial localization of lesions. To bridge this gap, Woo et al. [16] proposed CBAM, which progressively integrates channel attention with spatial attention. In brain tumor MRI segmentation, this enables the network to mitigate interference from normal tissue and concentrate specifically on tumor locations, facilitating the dual adaptive learning of feature selection and spatial attention.

Further advancements include Attention U-Net [12], which incorporates attention gating with skip connections. In pancreatic CT segmentation tasks, this efficiently mitigates abdominal background noise and attains enhanced localization precision for target structures with intricate morphology and indistinct borders. Similarly, Roy et al. [21] introduced the scSE module, which concurrently recalibrates features across both spatial and channel dimensions. Their work underscores the synergistic advantages of dual-dimensional attention across diverse benchmarks, including prostate and brain structural MRI segmentation. Cheng et al. [22] subsequently presented ResGANet, utilizing a residual group attention mechanism to concurrently model feature responses, demonstrating the efficacy of combining residual learning with attention strategies.

Recent research has also integrated Transformers to capture long-range dependencies. Chen et al. [23] introduced TransAttUNet, which integrates Transformer self-attention and global spatial attention within a U-shaped architecture, consistently outperforming existing methods in cross-modal tasks such as MRI and multi-organ CT segmentation. AlQurri et al. [24] further enhanced the U-Net architecture by incorporating three-layer attention modules with the Transformer, leveraging the interplay between decoder-side attention and dense skip connections. Moreover, Hussain et al. [25] introduced DCSSGA-UNet, a triadic mechanism encompassing channel, spatial, and semantic-guided attention to optimize multi-class biomedical image segmentation. In the specific context of breast cancer diagnosis, Xi et al. [26] demonstrated the efficacy of Transformer-based external attention, while Li et al. [27] utilized the hybrid architecture to confirm the robustness of Transformer-convolutional feature fusion for histopathology images.

Despite these advancements, current methodologies exhibit significant limitations: (1) standard channel attention often inadequately localizes fine-grained tumor boundaries, (2) conventional spatial attention relies on fixed receptive fields, rendering it unable to accommodate the high variability in tumor size and morphology, and (3) the simple

linear superposition of channel and spatial attention fails to fully exploit their potential synergy. These shortcomings provide the fundamental motivation for the Hybrid Multi-Scale Attention (HMSA) module developed in this study.

### 1.1.3. Medical Image Segmentation Based on Multiscale Feature Extraction

Multiscale feature extraction is a fundamental strategy for addressing target scale variability in medical imaging. The DeepLab series of networks, presented by Chen et al. [28], incorporated Atrous Spatial Pyramid Pooling (ASPP). This method enhances whole-heart segmentation in cardiac CT and MRI by concurrently delineating the intricate contours of cardiac chambers and the larger structures of adjacent major vessels via parallel atrous convolutions, thereby markedly improving the consistency of multiscale object segmentation. Similarly, Yu et al. [29] systematically integrated dilated convolutions into dense prediction tasks. By overlaying convolutional kernels with varying dilation rates to facilitate multi-scale context aggregation, they preserved high recognition accuracy for both capillaries and large arteries in retinal vascular segmentation. To further refine spatial context, Zhao et al. [30] introduced the Pyramid Pooling Module (PPM). In liver CT segmentation tasks, the PPM captures both local edge details and global spatial relationships via multiscale parallel pooling, effectively minimizing inter-organ missegmentation. In the domain of digital pathology, Schmitz et al. [31] employed a multi-scale fully convolutional network for tissue pathology image segmentation, addressing various levels from cell nucleus anomaly detection to comprehensive tissue structure analysis, thereby illustrating the distinct advantages of multiscale techniques in histopathological analysis.

Building on these foundations, recent research has pivoted toward dynamic fusion techniques for multi-scale features. Srivastava et al. [32] introduced MSRF-Net, which utilizes a Dual-Scale Dense Fusion (DSDF) block to facilitate bidirectional transfer of high- and low-resolution features. This approach has demonstrated robust generalization across benchmarks for colorectal polyp, skin lesion, and cell nuclei. Zhang et al. [33] presented SCIF-Net, which integrates a feature pyramid encoder with a multi-scale dynamic aggregation module to tackle substantial scale variations and low-contrast backgrounds in retinal and glandular segmentation. Furthermore, Yuan et al. [34] proposed a U-Net-based context fusion technique that assigns adaptive scale weights to each encoding step via a three-branch feature fusion (TBSFF) block, achieving precise segmentation of colon glands with minimal computational overhead.

Despite these advancements, existing multiscale methodologies are often constrained by static receptive fields and a lack of adaptive cross-scale integration, which hinders the capture of minute pathological features and the alignment of global semantics with fine boundaries. Furthermore, the decoupling of multiscale extraction and attention mechanisms limits the consistent enhancement of discriminative features. To address these deficiencies, we propose the Hybrid Multi-Scale Attention (HMSA) module, which synergistically integrates multiscale convolutions, dilated kernels, and adaptive attention within a unified framework.

## 2. Materials and Methods

### 2.1. Overall Model Architecture

This research introduces HT-Net, a novel histopathological image segmentation network designed with a dual-branch parallel encoder architecture. HT-Net is specifically engineered to address three critical challenges in breast cancer pathology: feature heterogeneity, class imbalance, and boundary ambiguity. The comprehensive framework is illustrated in Figure 1. The network processes  $256 \times 256 \times 3$  RGB histopathological image patches as input and generates a  $256 \times 256 \times 2$  segmentation map (differentiating tumor

from background). The architecture follows a three-stage encoder-fusion-decoder paradigm:

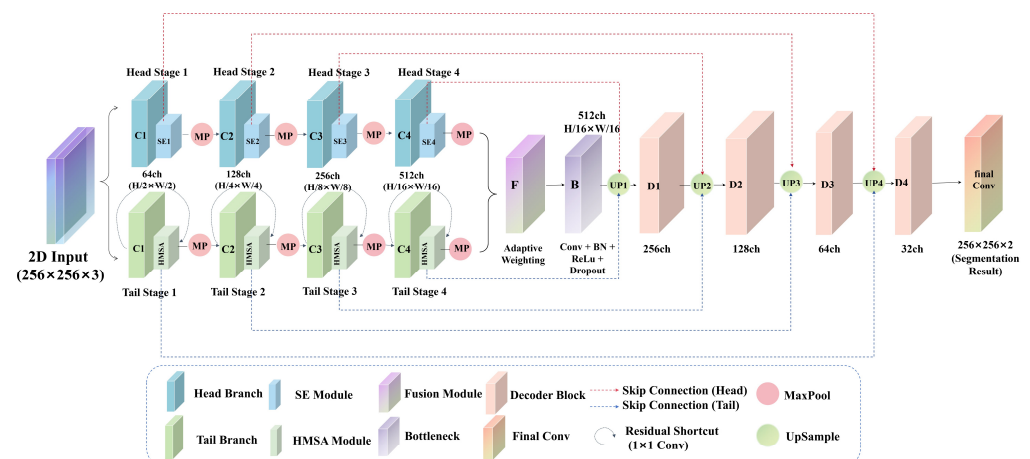


**Figure 1.** The Architecture of HT-Net.

1. **Dual-Branch Encoder:** The encoder features a parallel architecture comprising a HeadBranch and a TailBranch. These branches are specialized to perform distinct feature extraction for the background and tumor regions, respectively. Subsequently, a fusion module integrates the bottleneck features from both branches using a spatially adaptive weighting strategy.
2. **Symmetrical Decoder:** The decoder employs a U-Net-style symmetric architecture, utilizing skip connections to concurrently aggregate multi-scale information from both encoder branches and incrementally restoring spatial resolution to produce the final pixel-level prediction.

## 2.2. Dual-Branch Encoder

The dual-branch encoder consists of the HeadBranch and the TailBranch. While both branches process the same input image, they are equipped with distinct attention strategies tailored to their respective functional roles. Each branch follows a four-stage encoding hierarchy with channel depths of 64, 128, 256, and 512, respectively, while spatial resolution is progressively reduced using max-pooling operations.

### 2.2.1. HeadBranch Encoder

The Head Branch integrates a Squeeze-and-Excitation (SE) module to emphasize adaptive feature selection within the channel dimension. Each of its four encoding stages consists of a convolutional block, an SE module, and a max-pooling layer. The convolutional block is composed of two  $3 \times 3$  convolutional layers, each followed by batch normalization and a ReLU activation function. The detailed architecture specifications of the HeadBranch Encoder are provided in Table 1.

**Table 1.** Detailed Architecture of the HeadBranch Encoder.

| Stage | Layer Details                                          | Output Channels |
|-------|--------------------------------------------------------|-----------------|
| 0     | [Conv3 × 3 + BN + ReLU] × 2, SE, MaxPool2 × 2          | 64              |
| 1     | [Conv3 × 3 + BN + ReLU] × 2, SE, MaxPool2 × 2          | 128             |
| 2     | [Conv3 × 3 + BN + ReLU] × 2, SE, MaxPool2 × 2          | 256             |
| 3     | [Conv3 × 3 + BN + ReLU] × 2, SE, AuxHead, MaxPool2 × 2 | 512             |

The SE adaptively recalibrates channel-wise feature responses by explicitly modeling inter-channel dependencies. Specifically, for an input feature map  $F \in R^{C \times H \times W}$ , the SE

module first compresses the spatial dimension into a channel descriptor  $z \in R^C$  via global average pooling, as shown in Equation (1):

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j) \quad (1)$$

where  $z_c$  denotes the  $c$ th element of the channel descriptor. Subsequently, two fully connected layers are used to learn the nonlinear relationships between channels, and the Sigmoid function is applied to generate the channel attention weights  $s \in R^C$ , as shown in Equation (2):

$$s = \sigma(W^2 \cdot \delta(W^1 \cdot z)) \quad (2)$$

where  $W_1 \in R^{\frac{C}{r} \times C}$  and  $W_2 \in R^{C \times \frac{C}{r}}$  are learnable parameters,  $\delta(\cdot)$  denotes the ReLU function,  $\sigma(\cdot)$  denotes the Sigmoid function, and  $r$  is the scaling factor (set to 16 in this study). The final output is the channel-weighted feature map, calculated as shown in (3):

$$\tilde{F} = s \odot F \quad (3)$$

where  $\odot$  denotes element-wise multiplication. The number of output channels in the four encoding stages is 64, 128, 256, and 512, respectively. Following the final encoding stage, an auxiliary segmentation head is added to supervise the background region; this head consists of two convolutional layers and outputs a single-channel background probability map.

### 2.2.2. TailBranch Encoder

Similarly, the TailBranch comprises four encoding stages, substituting the SE module with the proposed Hybrid Multi-Scale Attention (HMSA) module. This module integrates a multi-scale channel attention branch with a multi-scale spatial attention branch to more effectively capture tumor border features, thereby enhancing spatial detail and multi-scale characteristics inherent in tumor regions. The detailed architecture of the HMSA module is shown in Figure 2.

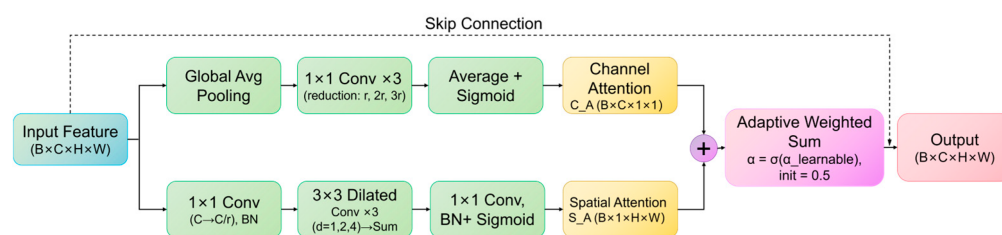


**Figure 2.** Architecture of the Hybrid Multi-Scale Attention (HMSA). Arrows indicate the direction of feature flow, while different colors distinguish the functional components of the module.

Specifically, the HMSA module consists of two parallel branches: a channel attention branch and a spatial attention branch, whose outputs are adaptively fused via a learnable weight.

The channel attention branch first applies Global Average Pooling to compress the input feature map from  $B \times C \times H \times W$  to  $B \times C \times 1 \times 1$ . The pooled features are then fed into three parallel bottleneck pathways with reduction ratios of  $r$ ,  $2r$ , and  $3r$  ( $r = 16$ ). The hidden width of the  $i$ -th pathway is  $\max(\lfloor \frac{C}{i \cdot r} \rfloor, C_{min})$ , where  $C_{min} = 8$  is a lower bound that prevents the bottleneck from collapsing at shallow stages (without it,  $\lfloor 64/48 \rfloor = 1$  at Stage 0). The resulting hidden widths are 8/8/8, 8/8/8, 16/8/8, and 32/16/10 for Stages 0–3, respectively. Consequently, the three pathways differ in width mainly at the deeper stages. Each pathway consists of two successive  $1 \times 1$  convolutions with an intermediate ReLU activation, where the first convolution reduces the channel dimension and the

second restores it to C. The outputs of the three pathways are averaged element-wise and passed through a Sigmoid activation function to generate channel attention weights of size  $B \times C \times 1 \times 1$ , as formulated in Equation (4):

$$CA = \sigma \left( \frac{1}{3} \sum_{i=1}^3 MLP_i(GAP(F)) \right) \quad (4)$$

where  $GAP(\cdot)$  denotes global average pooling, and  $MLP_i$  denotes the  $i$ -th multi-layer perceptron whose hidden width is  $\max(\lfloor \frac{C}{i \cdot r} \rfloor, C_{min})$ . The outputs of the three pathways are averaged before the Sigmoid activation  $\sigma$  is applied once.

The spatial attention branch first reduces the channel dimension from C to  $C/r$  via a  $1 \times 1$  convolution followed by Batch Normalization and ReLU activation. The reduced features are fed into three parallel  $3 \times 3$  dilated convolutions with dilation rates of 1, 2, and 4, corresponding to effective receptive fields of  $3 \times 3$ ,  $5 \times 5$ , and  $9 \times 9$ , respectively. The outputs are summed element-wise, followed by a  $1 \times 1$  convolution reducing channels to 1, Batch Normalization, and Sigmoid activation, producing spatial attention weights of size  $B \times 1 \times H \times W$ , as shown in Equations (5) and (6):

$$F_s = Conv_{1 \times 1}(F) \quad (5)$$

$$SA = \sigma(Conv_{1 \times 1}(\sum_{d \in \{1,2,4\}} DilatedConv_{\{3 \times 3, d\}}(F_s))) \quad (6)$$

where  $DilatedConv_{\{3 \times 3, d\}}(\cdot)$  denotes a  $3 \times 3$  dilated convolution with dilation rate  $d$ .

To adaptively balance the contributions of the two branches, a learnable scalar parameter  $\alpha$  is introduced and constrained to (0,1) via Sigmoid. The final output is obtained by element-wise multiplication of the input feature  $x$  with the fused attention weights ( $\alpha \cdot CA + (1 - \alpha) \cdot SA$ ), as formulated in Equations (7) and (8):

$$HMSA(F) = F \odot (\alpha \cdot CA + (1 - \alpha) \cdot SA) \quad (7)$$

$$\alpha = \sigma(\alpha_0) \quad (8)$$

where  $\alpha_0$  is a learnable scalar initialized to 0.5, so that the initial fusion weight is  $\alpha = \sigma(0.5) \approx 0.62$  (a slight initial preference for the channel attention branch);  $\alpha_0$  is then learned during training. The Tail branch incorporates auxiliary segmentation heads at each encoding stage to provide multi-scale deep supervision for tumor regions.

Furthermore, residual skip connections are implemented at each stage to facilitate gradient flow, utilizing  $1 \times 1$  convolution projection in deep architectures. The detailed architecture of the TailBranch is presented in Table 2.

**Table 2.** Encoder Structure of the TailBranch.

| Stage | Layer Details                                                            | Output Channels |
|-------|--------------------------------------------------------------------------|-----------------|
| 0     | [Conv $3 \times 3$ + BN + ReLU] $\times$ 2, HMSA, Skip, AuxHead, MaxPool | 64              |
| 1     | [Conv $3 \times 3$ + BN + ReLU] $\times$ 2, HMSA, Skip, AuxHead, MaxPool | 128             |
| 2     | [Conv $3 \times 3$ + BN + ReLU] $\times$ 2, HMSA, Skip, AuxHead, MaxPool | 256             |
| 3     | [Conv $3 \times 3$ + BN + ReLU] $\times$ 2, HMSA, Skip, AuxHead, MaxPool | 512             |

### 2.3. Dual-Branch Feature Fusion Module

The dual-branch feature fusion module integrates the bottleneck features (512 channels,  $H/16 \times W/16$ ) extracted from both the HeadBranch and TailBranch. This study proposes a spatially adaptive fusion technique that enables the network to dynamically adjust the contribution weights of the two branches according to local content. This mechanism ensures that features from the background-focused and tumor-focused branches are integrated contextually. To further refine the fused representation and improve generalization, the module incorporates a Convolution-BN-ReLU-Dropout architecture, which stabilizes the feature distribution and prevents overfitting during training.

First, the features from the two branches are projected to the same dimension via  $1 \times 1$  convolutions, as shown in Equations (9) and (10):

$$F_{head}^{proj} = Conv_{1 \times 1}(F_{head}) \quad (9)$$

$$F_{tail}^{proj} = Conv_{1 \times 1}(F_{tail}) \quad (10)$$

Next, the projected features are concatenated, and a lightweight network predicts the spatial fusion weights, as shown in Equation (11):

$$W = Softmax\left(Conv_{1 \times 1}\left(\delta\left(Conv_{1 \times 1}\left(\left[F_{head}^{proj}; F_{tail}^{proj}\right]\right)\right)\right)\right) \quad (11)$$

where  $[\cdot]$  denotes the concatenation operation, and  $W \in R^{2 \times H \times W}$  is a normalized two-channel weight map. The equation for the fused features is shown in (12):

$$F_{fused} = W_0 \odot F_{head}^{proj} + W_1 \odot F_{tail}^{proj} \quad (12)$$

where  $W_0$  and  $W_1$  represent the first and second channels of  $W$ , respectively. The fused features are further refined through convolutional blocks before being fed into the decoder.

### 2.4. U-Net Decoder

The decoder utilizes a U-Net-style symmetric architecture consisting of four distinct stages. At each stage, the feature map resolution is first doubled via transposed convolutions. These upsampled features are then concatenated with the corresponding feature maps from both the HeadBranch and TailBranch via skip connections. This multi-branch integration allows the decoder to recover spatial details while leveraging the specialized features extracted by the dual-branch encoder.

The channel depth is progressively reduced across the four stages to 256, 128, 64, and 32, respectively. Finally, a  $1 \times 1$  convolution layer projects these features into a 2-channel segmentation map (representing background and tumor classes). The decoding process at each stage is formally defined in Equation (13):

$$D_i = ConvBlock\left(\left[UpConv(D_{i+1}); F_i^{head}; F_i^{tail}\right]\right) \quad (13)$$

where  $UpConv(\cdot)$  denotes a  $2 \times$  transposed convolution,  $ConvBlock(\cdot)$  represents the convolution block, and  $F_i^{head}$  and  $F_i^{tail}$  represent the features from the  $i$ -th stage in the Headbranch and Tailbranch, respectively. The number of output channels for the four decoding stages is 256, 128, 64, and 32, respectively. The detailed architecture of the decoder is shown in Table 3.

**Table 3.** Architecture of the U-Net Decoder.

| Stage  | Layer Details                                                                         | Output Channels |
|--------|---------------------------------------------------------------------------------------|-----------------|
| 0      | TransConv2 $\times$ 2, Concat(Head + Tail), [Conv3 $\times$ 3 + BN + ReLU] $\times$ 2 | 256             |
| 1      | TransConv2 $\times$ 2, Concat(Head + Tail), [Conv3 $\times$ 3 + BN + ReLU] $\times$ 2 | 128             |
| 2      | TransConv2 $\times$ 2, Concat(Head + Tail), [Conv3 $\times$ 3 + BN + ReLU] $\times$ 2 | 64              |
| 3      | TransConv2 $\times$ 2, Concat(Head + Tail), [Conv3 $\times$ 3 + BN + ReLU] $\times$ 2 | 32              |
| Output | Conv1 $\times$ 1                                                                      | 2               |

### 2.5. Division-Based Supervision Loss Function

Deep supervision offers supplementary gradient signals to shallow layers through intermediate classifiers, hence mitigating the vanishing gradient issue in deep networks. Originally proposed by Lee et al. [35], this concept has been widely adopted in medical image segmentation. For instance, Dou et al. [36] integrated deep supervision into 3D medical imaging, while Zhou et al. [11] employed it in UNet++ to facilitate collaborative learning of multi-scale features.

In this study, we develop a task-division supervision loss function tailored to the unique characteristics of our dual-branch architecture. By delivering specific supervisory signals, this function directs the HeadBranch to focus on background recognition and the TailBranch to prioritize tumor detection, thereby fostering effective functional specialization and collaboration. The total objective function, HybridLoss, consists of a primary segmentation loss and task-specific auxiliary losses.

#### 2.5.1. Main Segmentation Loss

To ensure a fair comparison with baseline models, the main segmentation loss uses weighted cross-entropy loss, as defined in Equation (14):

$$L_{main} = -\sum_{c \in \{0,1\}} w_c \sum_{i,j} p_{c,i} \log(p_{c,i}) \quad (14)$$

where  $w_c$  represents class weights. We set  $w_0 = 1.0$  for the background, and  $w_1 = 5.0$  for the tumor class to mitigate the impact of class imbalance.

#### 2.5.2. Head Auxiliary Loss

The Head auxiliary loss is designed to steer the HeadBranch toward the accurate representation of background regions. We compute the binary cross-entropy over all pixels of the auxiliary output, using the background mask as the target, as shown in Equation (15):

$$L_{head} = \frac{1}{N} \sum_{i,j} BCE(P_{head}^{(i,j)}, M_{bg}^{(i,j)}) \quad (15)$$

where  $M_{bg} = 1$  [target < 0.5] denotes the background mask (1 for background, 0 for tumor), and  $N$  is the number of pixels of the auxiliary output. Because the loss is evaluated on both background and tumor pixels, a constant output cannot drive it to zero; reducing it below the level of a constant prediction requires the HeadBranch to spatially discriminate background from tumor.

#### 2.5.3. Tail Auxiliary Loss

The Tail auxiliary loss employs a multiscale deep supervision strategy to guide the TailBranch in learning tumor region features at each encoding stage. An auxiliary head is

attached to the output of each TailBranch stage (before max pooling), and the loss weights  $w_i$  are assigned as follows: Stage 0 (output resolution  $H$ ,  $w_0 = 0.4$ ), Stage 1 ( $H/2$ ,  $w_1 = 0.3$ ), Stage 2 ( $H/4$ ,  $w_2$ ), and Stage 3 ( $H/8$ ,  $w_3 = 0.1$ ), where  $H$  denotes the input resolution (i.e., 256, 128, 64, and 32 pixels for a  $256 \times 256$  input). Larger weights are thus given to the shallow, high-resolution stages to emphasize boundary details. No auxiliary head is attached to the  $H/16$  bottleneck. The loss for each scale is calculated by Equation (16):

$$L_{tail}^i = w_i \cdot \frac{\sum_{j,k} BCE\left(p_{tail}^{i,(j,k)}, M_{tumor}^{(j,k)}\right) \cdot W_{map}^{(j,k)}}{\sum_{j,k} W_{map}^{(j,k)}} \quad (16)$$

where  $i = 0, \dots, 3$  denotes the stage index, and where  $M_{tumor}^{(j,k)}$  is the binary label of the true tumor mask downsampled to the resolution of stage  $i$  by max pooling, so that a coarse cell is labeled as tumor whenever it contains at least one tumor pixel. Interpolation-based downsampling (bilinear or nearest-neighbor) of a binary mask is unreliable for small lesions: bilinear interpolation dilutes a lesion into fractional values, and nearest-neighbor sampling can skip it entirely. Max pooling avoids this, at the cost of slightly enlarging the target at coarse scales, which is acceptable because these heads serve only as auxiliary supervision and the coarsest scale has the smallest weight.  $W_{map}$  is the spatial weight map used to emphasize the learning of tumor regions, as shown in Equation (17):

$$W_{map}^{(j,k)} = \begin{cases} 5.0 & \text{if } M_{tumor}^{(j,k)} = 1 \\ 0.5 & \text{if } M_{tumor}^{(j,k)} = 0 \end{cases} \quad (17)$$

#### 2.5.4. Total Loss Function

The total loss function integrates these components to fully leverage the dual-branch architecture while maintaining comparability with existing methods, as shown in (18):

$$L_{total} = L_{main} + \lambda_{head} \cdot L_{head} + \lambda_{tail} \cdot \sum_{i=0}^3 L_{tail}^i \quad (18)$$

where  $\lambda_{head} = 0.3$ , and  $\lambda_{tail} = 0.4$ . This hybrid loss design ensures that HT-Net benefits from specialized branch supervision without deviating from standard performance benchmarks.

## 2.6. Datasets

### 2.6.1. Camelyon16 Dataset

The Camelyon16 dataset [37] is a widely recognized benchmark for histopathological image analysis. Co-released by Radboud University Medical Center and the University Medical Center Utrecht in the Netherlands, it is designed to assess algorithms for segmenting lymph node metastases in breast cancer. The dataset consists of 400 H&E-stained whole-slide images (WSIs), comprising 270 slides in the training set (160 normal slides and 110 tumor-infected slides) and 130 slides in the test set. Each tumor-positive slide is provided with ground truth annotations in XML format, delineating metastatic regions as defined by expert pathologists. This study utilizes this dataset as the principal evaluation benchmark.

### 2.6.2. Glas Dataset

The Glas (Gland Segmentation in Colon Histology Images Challenge Contest) dataset [38] serves as a definitive standard in colon gland segmentation, collaboratively published by Sirinukunwattana et al. It consists of 165 microscopic images obtained from 16 H&E-stained colon tissue slices. The data is partitioned into a training set of 85 images (comprising both benign and malignant glands) and a test set of 80 images (60 in TestA

and 20 in TestB). The dataset is characterized by diverse glandular morphologies, irregular boundaries, and complex lumen structures, often with neighboring glands in close proximity. We employ the Glas dataset as a secondary evaluation benchmark to assess the generalization capability of HT-Net in the context of colon glandular segmentation.

### 2.6.3. Data Preprocessing Method

Given that WSIs typically contain billions of pixels, they cannot be processed directly by deep learning models. This study utilizes a sliding window approach to extract fixed-size patches followed by rigorous quality control and mask generation. The specific steps are detailed as follows:

**Step 1. Patch Extraction:** The Level 0 pixel size of the slides is approximately  $0.243\ \mu\text{m}$  (Radboud University Medical Center) and  $0.226\ \mu\text{m}$  (University Medical Center Utrecht), equivalent to about  $40\times$  magnification. Patches were extracted at Level 1 of the pyramid, which is downsampled by a factor of 4 from Level 0 (approximately  $10\times$ ,  $0.90\text{--}0.97\ \mu\text{m}/\text{pixel}$ ), using non-overlapping sliding windows with a stride of 256 pixels, so that each  $256 \times 256$  patch covers about  $0.23\text{--}0.25\ \text{mm}$  of tissue. To filter out non-informative areas, a tissue threshold of 70% was applied; only patches containing at least 70% tissue were retained.

**Step 2. Quality Control:** To ensure the integrity of the training data, we performed a multi-dimensional screening process based on six criteria. (1) Patches with a pixel standard deviation below 15 were removed as low-contrast or solid-color; (2) Underexposed (mean brightness  $< 30$ ) or overexposed (mean brightness  $> 245$ ) patches were excluded; (3) Patches with less than 25% tissue area (mostly background) were discarded; (4) Blurry or out-of-focus images with a Laplacian variance below 80 were eliminated; (5) Faded or grayscale-like patches with a saturation score below 10 were removed; (6) WSI edge artifacts containing black regions exceeding 15% were discarded. Significantly, image patches with tumor annotations were preserved regardless of these thresholds to prevent the loss of critical positive samples. In total, approximately 65% of the patches extracted in Step 1 were discarded by the above screening, and the remaining 140,040 patches were used to build the final dataset.

**Step 3. Dataset Partitioning:** A stratified random splitting approach was employed to partition the data into training, validation, and test sets in a 6:2:2 ratio, maintaining uniform positive-to-negative sample ratios across all subsets. This resulted in a final dataset of 140,040 patches (84,024 for training, 28,008 for validation, and 28,008 for testing).

**Step 4. Segmentation Mask Generation:** Binary masks were generated for each patch based on the polygon coordinates in the XML annotation. Vertex coordinates were first converted from the Level 0 to Level 1. Subsequently, we utilized OpenCV's (version 4.8.1) `fillPoly` function to create  $256 \times 256$  binary masks, in which tumor regions were assigned a pixel value of 255 and background region a value of 0.

## 2.7. Experimental Setup

### 2.7.1. Evaluation Metrics

Segmentation performance was evaluated with four standard pixel-wise metrics: mean Intersection over Union (mIoU), mean Dice coefficient (mDice), mean Recall (mRecall), and mean Precision (mPrecision), whose definitions follow common practice in medical image segmentation [39]. For each class, the metrics are computed from the pixel-level true positives (TP), false positives (FP), and false negatives (FN), and then averaged over the  $K = 2$  classes (background and tumor for Camelyon16; background and gland for GlaS).

### 2.7.2. Implementation Details

To validate the efficacy of the proposed HT-Net and ensure a fair comparison with state-of-the-art methodologies, all models were trained under uniform experimental conditions. The primary objective function utilized was a weighted cross-entropy loss, with a weight of 5.0 assigned to the tumor class to address sample imbalance.

Network optimization was performed using the AdamW optimizer, with hyperparameters set to  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a weight decay of  $1 \times 10^{-3}$ . The initial learning rate was established at  $1 \times 10^{-4}$ , employing a cosine-smoothed annealing technique for adjustment, with a minimum learning rate of  $1 \times 10^{-6}$ . For the Camelyon16 dataset, the batch size was set to 16 with a maximum of 100 epochs. To ensure training stability, gradient clipping (maximum norm of 1.0) was applied, and an early stopping criterion was implemented to terminate training after 15 epochs without enhancement. Automatic Mixed Precision (AMP) was enabled to accelerate training and optimize GPU memory utilization. For the Glas dataset, the batch size was adjusted to 8 due to the reduced image dimensions, the maximum epochs were increased to 400, and the early stopping criterion was established to terminate training after 50 epochs without improvement.

Prior to being fed into the network, all images were scaled to  $256 \times 256$  pixels (Camelyon16 patches were extracted at this size and required no resizing; GlaS images were resized from their original size of approximately  $775 \times 522$  pixels). The input size of  $256 \times 256$  was chosen as a compromise between the field of view (about 0.25 mm at Level 1 for Camelyon16), computational cost, and GPU memory, and because it is divisible by the  $2^4$  downsampling required by the four-stage encoder. It was not tuned as a hyperparameter, and the sensitivity of HT-Net to patch size and resolution was not evaluated. Data augmentation strategies were employed to improve model robustness, including random rotation (angles randomly chosen from  $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ ), random horizontal and vertical flipping with a chance of 0.5, and random color jittering (with a variation range of 0.5 for brightness, contrast, and saturation, and 0.05 for hue). All input patches were normalized utilizing ImageNet pre-training parameters (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]).

For each model, the best epoch on each validation fold was selected, and results are reported as the mean  $\pm$  standard deviation over the five folds, using identical fold assignments for all models. Because the training sets of different folds overlap, fold-level scores are not independent; therefore, formal significance tests were not performed. All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) with 24 GB VRAM. Computational complexity (parameters, GFLOPs, inference latency, and GPU memory) was measured on the same GPU with a  $256 \times 256$  input for all models.

## 3. Results

### 3.1. Results on the Camelyon16 Dataset

To rigorously evaluate the segmentation performance of HT-Net, we selected five state-of-the-art approaches as comparative baselines. These models represent the most prominent technical paradigms in current medical image segmentation:

1. U-Net [10] and UNet++ [11]: These serve as the primary convolutional benchmarks. U-Net establishes the baseline for standard encoder–decoder architectures, while UNet++ represents the advancement of nested dense connections.
2. UTransNet [40]: This model incorporates a channel Transformer with the skip connections, representing a significant evolution attention-augmented U-shaped networks.

3. VM-UNet [41]: Based on the Visual State Space Model (Mamba), this framework illustrates the recent application of innovative sequence modeling paradigms in the medical imaging domain.
4. nnWNet [42]: Integrating the WNet architecture into the nnUNet automated training framework, this model facilitates continuous feature propagation through the parallel design of local convolutional branches and global Transformer branches. It is currently one of the most competitive methods in terms of overall performance.

By encompassing pure convolutional networks, attention-enhanced models, state-space models, and hybrid Transformer-CNN architectures, these five methodologies provide a multidimensional benchmark. This selection ensures a comprehensive assessment of HT-Net's efficacy against the diverse technical strategies currently dominating the field.

Table 4 displays the quantitative comparison results between HT-Net and state-of-the-art methodologies on the Camelyon16 dataset. HT-Net achieved superior performance across all primary metrics, with a mIoU of 82.05% and a mDice of 89.37%, thereby substantiating the efficacy of the proposed architecture.

**Table 4.** Comparison of Segmentation Performance Between HT-Net and Baseline Methods on the Camelyon16 Dataset.

| Model         | mDice               | mIoU                | mRecall             | mPrecision          |
|---------------|---------------------|---------------------|---------------------|---------------------|
| U-Net         | 87.09 ± 1.86        | 79.28 ± 2.31        | 85.35 ± 2.14        | 89.44 ± 2.03        |
| UNet++        | 86.88 ± 1.94        | 78.98 ± 2.38        | 85.69 ± 2.21        | 88.87 ± 2.16        |
| UCTransNet    | 87.75 ± 1.72        | 79.93 ± 2.16        | 86.84 ± 1.96        | 88.73 ± 2.05        |
| VM-UNet       | 88.07 ± 1.63        | 80.11 ± 2.08        | 87.22 ± 1.87        | 89.03 ± 1.94        |
| nnWNet        | 88.51 ± 1.52        | 80.97 ± 1.96        | 88.27 ± 1.78        | 89.16 ± 1.86        |
| HT-Net (ours) | <b>89.37 ± 1.18</b> | <b>82.05 ± 1.67</b> | <b>89.54 ± 1.73</b> | <b>89.40 ± 1.50</b> |

Note: Bold values indicate the best performance among all compared methods.

HT-Net outperformed classical convolutional networks, demonstrating improvements of 2.30 and 2.49 percentage points in mDice relative to U-Net and UNet++, respectively. Correspondingly, it showed gains of 2.77 and 3.07 percentage points in mIoU. These results suggest that the dual-branch, “division-of-labor” strategy provides a significant advantage in distinguishing between complex tumor features and background characteristics.

Compared to the attention-augmented UCTransNet [40], HT-Net showed enhancements of 1.62 and 2.12 percentage points in mDice and mIoU, respectively. This indicates that the dual-branch differential attention mechanism possesses superior regional discrimination capabilities compared to a single-branch channel Transformer. Furthermore, HT-Net surpassed the state-space-based VM-UNet [41] by 1.30 and 1.94 percentage points; the multiscale dilated convolutions within the HMSA module effectively address the inherent limitations of SSM-based approaches in capturing local boundary details.

Finally, in comparison to nnWNet [42], which showed the most competitive baseline performance, HT-Net maintained a lead of 0.86 and 1.08 percentage points in mDice and mIoU. The implementation of a task-specific supervision loss function further refines the functional specialization of the dual branches, allowing the model to sustain high detection sensitivity for tumorous regions while preserving background integrity.

Figure 3 provides a visual comparison of segmentation results between HT-Net and baseline methods on the Camelyon16 test set. Each row depicts a challenging histopathological image patch, while columns represent, from left to right: the input image, the prediction of baseline methods, HT-Net prediction, and the ground truth (GT). Overall, HT-Net's predictions outperform the baseline models in both morphological integrity and boundary precision.

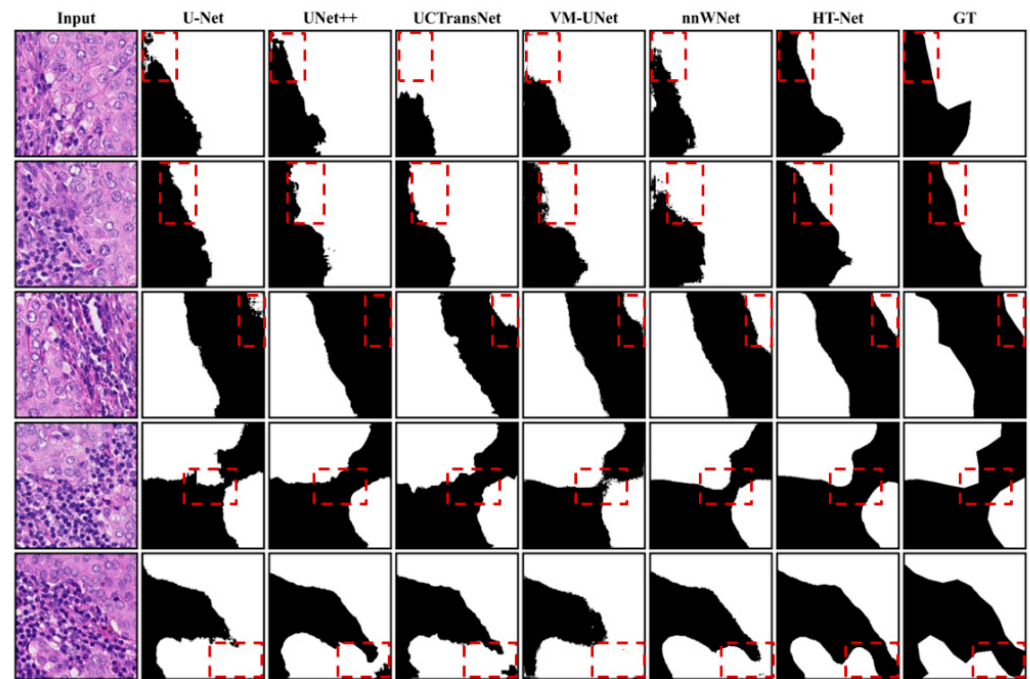


**Figure 3.** Visual Comparison of segmentation results between HT-Net and baseline methods on the Camelyon16 test set. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. GT represents the ground truth.

In instances involving large tumorous regions (Rows 1 and 2), both U-Net and UNet++ exhibit varying degrees of under-segmentation characterized by irregular edges. In contrast, HT-Net effectively preserves the primary tumor morphology, yielding smoother boundaries that closely align with the ground truth. In samples with distinct tumor-background delineations (Rows 3 and 4), while all three methodologies correctly identify the core tumor areas, U-Net and UNet++ show significant misclassifications and boundary artifacts. By leveraging the multiscale spatial attention features of its HMSA module, HT-Net achieves superior precision and clarity in delineating these transition zones.

For the most difficult cases, where clinical characteristics are intricate and local textures closely resemble tumor tissue (Rows 5 and 6), U-Net produces a substantial amount of false-positive background noise. Although UNet++ offers some improvement, isolated artifacts remain evident. Conversely, HT-Net effectively mitigates background interference through its task-specific supervision strategy, maintaining a robust focus on actual tumor locations. These visualization findings corroborate the quantitative metrics, further affirming the superiority of HT-Net in histopathological tumor segmentation.

### 3.2. Results on the Glas Dataset

We employed five same baseline methods to further assess the generalization capabilities of HT-Net on the Glas colon gland segmentation dataset.

The quantitative results conclusively demonstrate the efficacy and generalization capabilities of the proposed strategy for the Glas colon gland segmentation challenge. Table 5 presents a performance comparison between HT-Net and leading state-of-the-art approaches on the Glas dataset. HT-Net achieved superior performance, reaching 93.04% mDice and 86.95 mIoU, thereby confirming its effectiveness in the context of colon adenocarcinoma segmentation.

**Table 5.** Comparison of segmentation performance between HT-Net and baseline methods on the Glas dataset.

| Model         | mDice               | mIoU                | mRecall             | mPrecision          |
|---------------|---------------------|---------------------|---------------------|---------------------|
| U-Net         | 92.19 ± 1.12        | 85.52 ± 1.54        | 92.22 ± 1.26        | 92.21 ± 0.98        |
| UNet++        | 91.66 ± 1.35        | 84.59 ± 1.61        | 91.60 ± 1.48        | 91.73 ± 1.17        |
| UCTransNet    | 92.03 ± 1.08        | 85.26 ± 1.56        | 92.06 ± 1.19        | 92.18 ± 1.02        |
| VM-UNet       | 92.51 ± 0.96        | 86.07 ± 1.38        | 92.43 ± 1.04        | 92.84 ± 0.87        |
| nnWNet        | 92.74 ± 0.89        | 86.48 ± 1.17        | 92.77 ± 0.95        | 92.73 ± 0.82        |
| HT-Net (ours) | <b>93.04 ± 0.75</b> | <b>86.95 ± 0.99</b> | <b>93.03 ± 0.81</b> | <b>93.09 ± 0.64</b> |

Note: Bold values indicate the best performance among all compared methods.

HT-Net outperformed classical convolutional networks, demonstrating gains of 0.85 and 1.38 percentage points in mDice, and 1.43 and 2.36 percentage points in mIoU compared to U-Net and UNet++, respectively. These results indicate that the dual-branch collaborative strategy provides a substantial advantage in reconciling the morphological disparities between glandular structures and stromal characteristics.

Compared to the attention-augmented UCTransNet [40], HT-Net showed enhancements of 1.01 and 1.69 percentage points in mDice and mIoU, respectively. This signifies that the dual-branch differential attention mechanism possesses superior regional discrimination capabilities compared to single-branch Transformer architectures. Furthermore, HT-Net surpassed the state-space-based VM-UNet [41] by 0.53 and 0.88 percentage points. The multiscale dilated convolutions within the HMSA module effectively address the limitations of SSM-based approaches in resolving fine features at glandular boundaries.

Finally, in comparison to nnWNet [42], which showed the most competitive performance, HT-Net maintained an advantage of 0.30 and 0.47 percentage points in mDice and mIoU. The task-specific supervision loss function further enhances the functional specialization of the dual branches, enabling the model to achieve higher detection accuracy for glandular regions while preserving the integrity of the stromal background.

Figure 4 presents a visual comparison of segmentation results between HT-Net and baseline methods on the Glas test set. Each row depicts a representative challenging image patch, while the columns display, from left to right: the input image, the prediction of baseline methods, HT-Net prediction, and the ground truth (GT). Overall, HT-Net's predictions outperform the baseline models in both morphological integrity and boundary precision.

In instances involving large glandular regions (Rows 1 and 2), both U-Net and UNet++ exhibit varying degrees of omission characterized by irregular edges. Conversely, HT-Net effectively preserves the primary glandular morphology, generating smoother boundaries that align closely with the ground truth. In samples with distinct tissue delineations (Rows 3 and 4), while all three methodologies successfully identify the core glandular areas, U-Net and UNet++ show significant misclassifications and under-segmentation at the boundary transitions. By leveraging the multiscale spatial attention features of its HMSA module, HT-Net achieves enhanced precision and clarity in delineating these complex boundary regions.

For difficult cases with intricate clinical features where local textures closely resemble glandular tissue (Rows 5 and 6), U-Net generates a substantial number of false-positive background artifacts. Although UNet++ offers relative improvements, isolated noise sites remain prevalent. In contrast, HT-Net efficiently mitigates stromal background interference through its task-specific supervision strategy, maintaining a robust focus on the actual target locations. These visualization findings corroborate the quantitative metrics, further affirming the superiority of HT-Net for segmentation tasks in histopathological imaging.

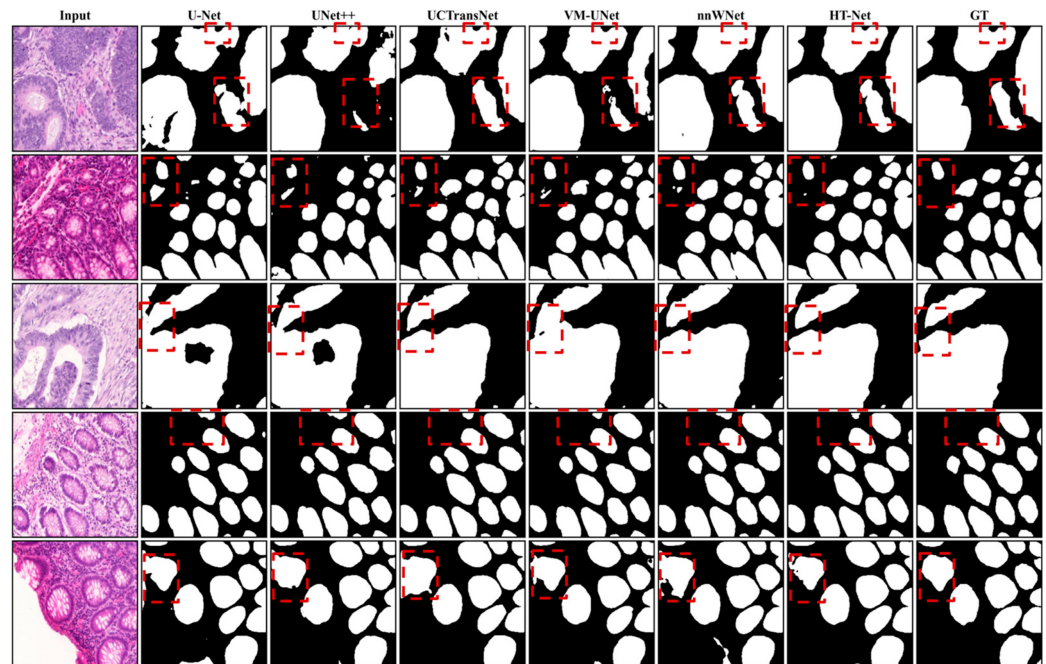


**Figure 4.** Visual Comparison of segmentation results between HT-Net and baseline methods on the Glas test set. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. GT represents the ground truth.

### 3.3. Ablation Study

We conducted progressive ablation experiments on the Camelyon16 dataset to evaluate the individual contributions of the dual-branch architecture, the HMSA module, and the task-specific supervision strategy to the overall performance of HT-Net. We systematically introduced each component and monitored performance variations to progressively validate their efficacy. The specific configurations are defined as follows:

1. **Baseline:** A conventional U-Net encoder–decoder architecture, devoid of attention mechanisms or auxiliary supervision. This model establishes a performance lower bound and illustrates the intrinsic constraints of the single-branch homogenous feature extraction methodology.
2. **Dual-SE:** The single-branch encoder is enhanced into a dual-branch parallel architecture, using SE attention modules in both the HeadBranch and TailBranch. This configuration aims to evaluate the performance enhancements intrinsic to the dual-branch parallel architecture—specifically, whether increasing the number of branches alleviates regional feature heterogeneity when the attention mechanism remains constant.
3. **Dual-SE-HMSA:** The SE module in the TailBranch is replaced by the proposed HMSA module, while the SE module in the HeadBranch remains unchanged. This variation independently assess the efficacy of the HMSA module, investigating if the incorporation of a differentiated attention mechanism can enhance the multiscale perception of tumor boundary characteristics.
4. **HT-Net (Comprehensive Model):** This configuration enhances the Dual-SE-HMSA by incorporating a task-specific supervision loss function, which utilizes background-region-targeted auxiliary supervision for the HeadBranch and multiscale tumor-region deep supervision for the TailBranch. This stage determines if the task-specific supervision technique can facilitate effective functional distinction between the two branches at the objective function level, hence optimizing overall segmentation performance.

The four configurations exhibit a clear evolutionary trend: single-branch baseline → dual-branch homogenous attention → dual-branch differentiated attention → dual-branch

differentiated attention with task-specific supervision. By altering only one variable at each stage, we ensured the validity of the conclusions derived from this ablation study. Each stage alters only one variable, hence guaranteeing the validity of the conclusions gained from the ablation experiment.

### 3.3.1. Ablation Results on the Camelyon16

Specifically, the quantitative and visual results of the ablation experiment for HT-Net components on the Camelyon16 dataset are shown in Table 6 and Figure 5, respectively. Table 6 presents the ablation study results on the Camelyon16 dataset, demonstrating the incremental performance gains achieved by each component of the proposed HT-Net. Compared to the baseline U-Net, the integration of the Dual-SE and HMSA modules consistently improved segmentation accuracy across all metrics. Ultimately, the full HT-Net achieved superior performance, reaching 89.39% mDice and 82.06% mIoU, thereby validating the effectiveness of the proposed architectural enhancements for histopathology image analysis.

**Table 6.** Ablation Study Results on the Camelyon16 Dataset.

| Model         | mDice               | mIoU                | mRecall             | mPrecision          |
|---------------|---------------------|---------------------|---------------------|---------------------|
| U-Net         | 87.09 ± 1.86        | 79.28 ± 2.31        | 85.35 ± 2.14        | 89.44 ± 2.03        |
| Dual-SE       | 87.93 ± 1.72        | 80.17 ± 2.17        | 87.19 ± 1.96        | 88.62 ± 1.91        |
| Dual-SE-HMSA  | 88.76 ± 1.43        | 81.53 ± 1.89        | 88.31 ± 1.82        | 89.28 ± 1.69        |
| HT-Net (ours) | <b>89.37 ± 1.18</b> | <b>82.05 ± 1.67</b> | <b>89.54 ± 1.73</b> | <b>89.40 ± 1.50</b> |

Note: Bold values indicate the best performance among all compared methods.

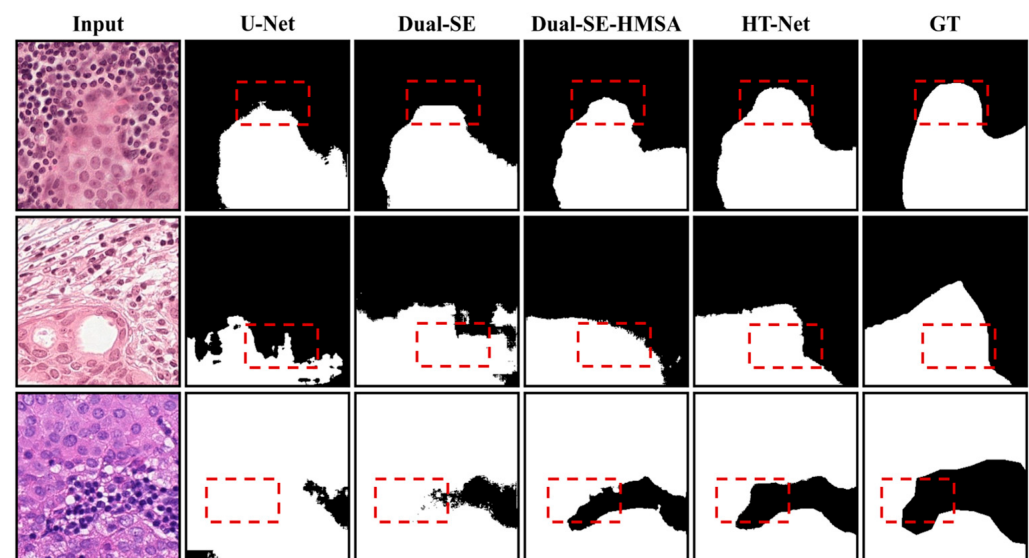


**Figure 5.** Visual Comparison of segmentation results from ablation experiments on various HT-Net components using the Camelyon16 dataset. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. GT represents the ground truth.

As illustrated in Figure 5, the Baseline (U-Net) produces considerable “jaggedness” in predicted boundaries, characterized by a high volume of misclassified pixels in the tumor-background transition zones and a failure to identify fine substructures. Following the implementation of the dual-branch architecture (Dual-SE), the continuity of the tumor body is enhanced, and background misclassification rates are reduced; however, irregular noise persists at the boundaries.

By replacing the SE module with the HMSA module (Dual-SE-HMSA), the multiscale dilated convolutions effectively expand the receptive field, yielding smoother and more

continuous boundary contours while better preserving substructure integrity. Finally, the incorporation of task-specific auxiliary supervision in the full HT-Net significantly improves background suppression. This configuration achieves the optimal overlap between the predicted tumor regions and the ground truth (GT). These results confirm that our functional specialization strategy—in which the HeadBranch focuses on background representation and the TailBranch prioritizes tumor detection—effectively directs the feature learning of the dual branches to attain peak segmentation performance.

### 3.3.2. Ablation Results on the Glas Dataset

Moreover, we performed the ablation study on the Glas dataset and the quantitative and visual results for HT-Net components are shown in Table 7 and Figure 6, respectively. Table 7 summarizes the ablation study results on the Glas dataset, illustrating the individual contributions of the proposed modules to the overall segmentation performance. The baseline U-Net is progressively enhanced by the addition of the Dual-SE and HMSA components, resulting in steady improvements across all evaluated metrics. The complete HT-Net architecture achieves the highest performance, with mDice and mIoU reaching 93.04% and 86.95%, respectively. These results further confirm the robustness and efficacy of the integrated modules in capturing complex morphological features in gland segmentation tasks.

**Table 7.** Ablation Study Results on the Glas Dataset.

| Model         | mDice               | mIoU                | mRecall             | mPrecision          |
|---------------|---------------------|---------------------|---------------------|---------------------|
| U-Net         | 92.19 ± 1.12        | 85.52 ± 1.54        | 92.22 ± 1.26        | 92.21 ± 0.98        |
| Dual-SE       | 92.30 ± 1.05        | 85.70 ± 1.76        | 92.50 ± 1.13        | 92.26 ± 0.94        |
| Dual-SE-HMSA  | 92.61 ± 0.92        | 86.24 ± 1.61        | 92.70 ± 0.98        | 92.55 ± 0.86        |
| HT-Net (ours) | <b>93.04 ± 0.75</b> | <b>86.95 ± 0.99</b> | <b>93.03 ± 0.81</b> | <b>93.09 ± 0.64</b> |

Note: Bold values indicate the best performance among all compared methods.

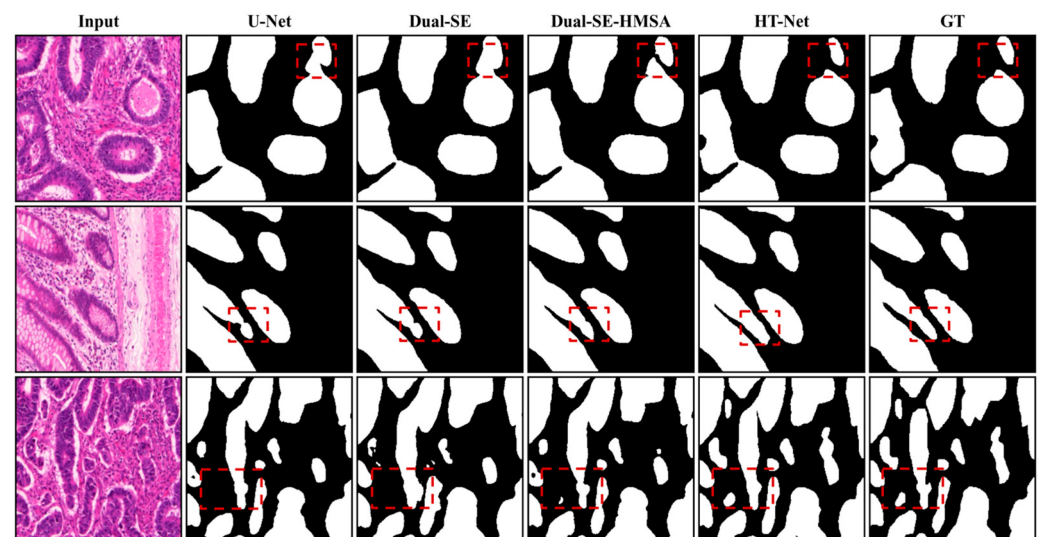


**Figure 6.** Visual Comparison of segmentation results from ablation experiments on various HT-Net components using the Glas dataset. Regions highlighted by the dashed boxes are for better visualization of the prediction differences. GT represents the ground truth.

The visualization results depicted in Figure 6 substantiate the quantitative findings: the Baseline (U-Net) generates relatively imprecise representations of glandular boundaries, exhibiting indistinct separations between the glandular lumen and the adjacent stroma. In

this configuration, several small glands are either overlooked or incorrectly predicted as merged entities.

Following the implementation of the dual-branch architecture (Dual-SE), the segmentation fidelity of the primary glandular regions is improved, and the incidence of inter-gland fusion is reduced. However, the gains in boundary precision remain limited, with instances of over-segmentation and under-segmentation persisting in complex areas. Replacing the SE module with the HMSA module (Dual-SE-HMSA) markedly improves the model's capacity to discern glandular contours. By incorporating multiscale receptive fields, this module yields more defined boundaries while better preserving internal glandular structures.

Finally, the introduction of task-specific auxiliary supervision in the full HT-Net further reduces misclassified regions within the stromal background. This approach achieves optimal alignment between the predicted glandular boundaries and the ground truth (GT), producing the most precise overall segmentation results. These findings further substantiate the efficacy of the task-specific supervision strategy in facilitating the functional specialization of dual branches when navigating the diverse morphological variations inherent in glandular segmentation.

#### 4. Discussion

Table 8 compares the computational complexity of all models under identical settings. HT-Net has the fewest parameters (24.0 M), and its computational cost (46.3 GFLOPs) is lower than that of U-Net, UNet++, and UCTransNet, but higher than that of nnWNet (18.9 GFLOPs). Its latency (11.4 ms) and GPU memory usage (2625.7 MB) are slightly higher than those of U-Net (8.9 ms, 2285.6 MB) and nnWNet (10.7 ms, 2518.3 MB), and clearly lower than those of UNet++ (18.5 ms, 2901.6 MB) and UCTransNet (15.6 ms, 3427.5 MB). The dual-branch design therefore adds some overhead relative to the lightest baselines, but the overhead is modest in absolute terms: HT-Net processes a  $256 \times 256$  image in 11.4 ms using about 2.6 GB of GPU memory on a single RTX 4090. However, these measurements were obtained on a workstation GPU, and its suitability for resource-constrained clinical deployments still needs to be verified on the target hardware.

HT-Net is intended as a decision-support tool for pathologists assessing sentinel lymph node metastasis, acting as a second reader rather than replacing the pathologist. Following the deployment-oriented descriptions of AI-assisted diagnostic workflows in medical imaging [43], we outline the envisioned pipeline, which has not yet been validated end to end. (1) A scanned whole-slide image (WSI) is loaded and a tissue mask is obtained from a low-resolution thumbnail to exclude the glass background. (2) The slide is tiled at pyramid Level 1 (approximately  $10\times$ ) into  $256 \times 256$  patches with a stride of 256, the same setting used to build our training data, and patches with less than 70% tissue are skipped. (3) The patches are processed in batches by HT-Net in inference mode, in which only the main segmentation output is computed and the auxiliary heads are not used. (4) The patch predictions are written back to their tile coordinates to form a slide-level mask, which can be mapped to Level 0 coordinates and shown in the slide viewer as a semi-transparent overlay or contour, leaving the final decision to the pathologist. Tile-boundary artifacts can be reduced by using overlapping tiles (e.g., a stride of 128) and averaging the overlapping predictions, at roughly four times the computation.

Regarding resources, HT-Net processes a  $256 \times 256$  patch in 11.4 ms and uses about 2.6 GB of GPU memory on a single RTX 4090 (Table 8). Based on this latency, the pure inference throughput is roughly 88 patches per second, so a slide with 5000 tissue patches would require on the order of one minute of network inference; this estimate excludes slide reading and mask stitching, which were not measured. Because the memory footprint

fits a single workstation-grade GPU, the model can be deployed locally next to the slide viewer, avoiding the transfer of gigapixel images to external servers, while a server-based deployment remains an option for batch processing of many slides. We emphasize that our experiments evaluate patch-level segmentation on pre-extracted patches. End-to-end slide-level evaluation, measurement of complete per-slide processing time, integration into clinical viewing software, and prospective validation with pathologists were not performed and are left for future work.

Furthermore, while validated on breast and colonic tissues, the model's generalizability to other pathological types warrants further investigation. Future research will focus on three primary directions: (1) Model Compression: Implementing knowledge distillation or network pruning to develop a lightweight version of HT-Net suitable for resource-constrained clinical settings without sacrificing accuracy. (2) Domain Expansion: Assessing the framework's versatility across a broader spectrum of histopathological datasets, with a specific focus on challenging scenarios such as diminutive glands and sparse metastatic distributions. (3) Hyperparameter Optimization: The hyperparameters of HT-Net, such as the auxiliary loss weights, the class weight and the learning rate, were set manually in this study. Automated optimization of these parameters, for example, with evolutionary algorithms such as genetic algorithms [44], is a promising direction for improving performance.

**Table 8.** Computational complexity of HT-Net and the compared methods. All models were measured under identical conditions with a  $256 \times 256$  input on an NVIDIA RTX 4090 GPU (batch size 32, FP32); latency is the average over 100 forward passes after 50 warm-up iterations. N/A GFLOPs of VM-UNet are not reported because its selective-scan operator is implemented as a custom CUDA kernel that the FLOPs profiler used for the other models ([thop/fvcore]) does not count, so the same procedure would give an incomplete value.

| Model         | Params (M) | GFLOPs | Latency (ms) | GPU Memory (MB) |
|---------------|------------|--------|--------------|-----------------|
| U-Net         | 31.0       | 54.7   | 8.9          | 2285.6          |
| UNet++        | 35.8       | 137.8  | 18.5         | 2901.6          |
| UCTransNet    | 66.5       | 50.8   | 15.6         | 3427.5          |
| VM-UNet       | 27.4       | N/A    | 9.3          | 2387.8          |
| nnWNet        | 28.4       | 18.9   | 10.7         | 2518.3          |
| HT-Net (ours) | 24.0       | 46.3   | 11.4         | 2625.7          |

## 5. Conclusions

This study introduced HT-Net, a novel dual-branch tumor segmentation network designed to address the critical challenges of feature heterogeneity, class imbalance, and indistinct boundaries in histopathological imaging. By integrating differentiated encoding, multiscale attention, and task-specific supervision, HT-Net achieves a synergistic balance between background suppression and tumor localization. The proposed architecture is conceptually aligned with biomimetic principles of functional specialization, selective attention, and complementary information processing, in which different processing pathways contribute specialized representations that are adaptively integrated for robust perception. Our key contributions are summarized as follows:

1. **Dual-Branch Parallel Encoding:** We developed a parallel architecture where the Head-Branch (utilizing SE attention) focuses on background feature selection, while the TailBranch (utilizing the proposed HMSA module) targets tumor-specific characteristics. Ablation results confirm that this collaborative feature extraction improves mDice and mIoU by 0.84 and 0.89 percentage points, respectively, over the single-branch baseline.

2. **Hybrid Multi-Scale Attention (HMSA):** The HMSA module effectively captures multi-granularity channel dependencies and expansive spatial contexts through parallel dilated convolutions (dilation rates: 1, 2, 4). This differentiated attention mechanism outperformed single-scale alternatives, enhancing mDice by 0.83 and mIoU by 1.36 percentage points.
3. **Task-Specific Supervision:** A specialized loss function was formulated to enforce functional differentiation. By applying background-targeted supervision to the HeadBranch and multiscale deep supervision to the TailBranch, we mitigated class imbalance and optimized inter-branch collaboration, yielding further gains of 0.61 (mDice) and 0.52 (mIoU) percentage points.
4. **SOTA Performance and Generalization:** Comprehensive evaluations on the Camelyon16 and Glas datasets demonstrate that HT-Net consistently outperforms leading methods, including nnWNet and VM-UNet. Specifically, HT-Net achieved a mDice of 89.37% on Camelyon16 and 93.04% on Glas, underscoring its robustness across different pathological contexts and morphological variations.

**Author Contributions:** Conceptualization, L.R. and L.H.; methodology, L.R. and H.C.; software, L.R. and H.C.; validation, L.R., H.C. and D.T.; formal analysis, L.R.; investigation, J.Y. and D.T.; resources, J.Y.; data curation, H.C. and J.Y.; writing—original draft preparation, L.R. and H.C.; writing—review and editing, J.Y., L.H. and D.T.; visualization, L.R.; supervision, L.H.; project administration, L.H.; funding acquisition, L.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Chengdu Municipal Science and Technology Program, grant number 2026-YF08-00060-GX, and the Sichuan Provincial Major Science and Technology Special Project: No. 2022ZDZX0001. The APC was funded by Chengdu University of Information Technology.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets analyzed in this study are publicly available. The Camelyon16 dataset can be accessed at <https://camelyon16.grand-challenge.org/> (accessed on 11 October 2026), and the Glas dataset at <https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest/> (accessed on 11 October 2026; or via the original publication: [38]). No new data were created in this study. The source code of HT-Net is not publicly available at this time but is available from the corresponding author upon reasonable request.

**Acknowledgments:** GPT/ChatGPT (GPT-4o, OpenAI) was used only for English translation and language polishing to improve the clarity, grammar, and readability of the manuscript. The scientific content, research design, methodology, results, analysis, and conclusions were developed and verified by the authors.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|        |                                      |
|--------|--------------------------------------|
| WSI    | Whole-Slide Image                    |
| SLN    | Sentinel Lymph Node                  |
| MIL    | Multiple Instance Learning           |
| SE     | Squeeze-and-Excitation               |
| HMSA   | Hybrid Multi-Scale Attention         |
| HT-Net | Head–Tail Network                    |
| CBAM   | Convolutional Block Attention Module |
| mDice  | mean Dice coefficient                |
| mIoU   | mean Intersection over Union         |

## References

1. Bray, F.; McCarron, P.; Parkin, D.M. The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res.* **2004**, *6*, 229–239. [\[CrossRef\]](#)
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [\[CrossRef\]](#)
3. Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; Beck, A.H. Deep learning for identifying metastatic breast cancer. *arXiv* **2016**, arXiv:1606.05718.
4. Subramanian, R.; Rubi, R.D.; Kasavaraju, A.K.; Jain, S.; Guptha, S.; Pingali, S.R. Automatic classification of sentinel lymph node (SLN) metastases in breast carcinoma whole slide image (WSI) through densenet deep learning network. *J. Diagn. Pathol. Oncol.* **2020**, *5*, 144–150. [\[CrossRef\]](#)
5. Bejnordi, B.E.; Veta, M.; Van Diest, P.J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J.A.W.M.; Hermesen, M.; Manson, Q.F.; Balkenhol, M.; et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **2017**, *318*, 2199–2210. [\[CrossRef\]](#)
6. Ding, K.; Zhou, M.; Wang, H.; Gevaert, O.; Metaxas, D.; Zhang, S. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. *Sci. Data* **2023**, *10*, 231. [\[CrossRef\]](#)
7. Siddique, N.; Paheding, S.; Elkin, C.P.; Devabhaktuni, V. U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **2021**, *9*, 82031–82057. [\[CrossRef\]](#)
8. Wang, Z.; Yu, L.; Ding, X.; Liao, X.; Wang, L. Lymph node metastasis prediction from whole slide images with transformer-guided multi-instance learning and knowledge transfer. *IEEE Trans. Med. Imaging* **2022**, *41*, 2777–2787. [\[CrossRef\]](#)
9. Cheplygina, V.; Sørensen, L.; Tax, D.M.; de Bruijne, M.; Loog, M. Label stability in multiple instance learning. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
11. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
12. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
13. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
14. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision Workshops*; Springer: Cham, Switzerland, 2022; pp. 205–218.
15. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
16. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8–14 September 2018; pp. 3–19.
17. Alom, M.Z.; Yakopcic, C.; Hasan, M.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* **2019**, *6*, 014006. [\[CrossRef\]](#)
18. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. UNet3+: A full-scale connected UNet for medical image segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, NY, USA, 2020; pp. 1055–1059.
19. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [\[CrossRef\]](#)
20. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [\[CrossRef\]](#)
21. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2018; pp. 421–429.
22. Cheng, J.; Tian, S.; Yu, L.; Gao, C.; Kang, X.; Ma, X.; Wu, W.; Liu, S.; Lu, H. Resganet: Residual group attention network for medical image classification and segmentation. *Med. Image Anal.* **2022**, *76*, 102313. [\[CrossRef\]](#)
23. Chen, B.; Liu, Y.; Zhang, Z.; Lu, G.; Kong, A.W.K. TransAttUNet: Multi-level attention-guided U-Net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 55–68. [\[CrossRef\]](#)
24. AlQurri, A.; Almekkawy, M. Improved UNet with attention for medical image segmentation. *Sensors* **2023**, *23*, 8589. [\[CrossRef\]](#)

25. Hussain, T.; Shouno, H.; Mohammed, M.A.; Marhoon, H.A.; Alam, T. DCSSGA-UNet: Biomedical image segmentation with DenseNet channel spatial and semantic guidance attention. *Knowl.-Based Syst.* **2025**, *314*, 113233. [\[CrossRef\]](#)
26. Xi, R.; Ahmed, A.; Zeng, X.; Hou, M. A novel transformers-based external attention framework for breast cancer diagnosis. *Biomed. Signal Process. Control* **2025**, *110*, 108065. [\[CrossRef\]](#)
27. Li, X.; Pang, S.; Zhang, R.; Zhu, J.; Fu, X.; Tian, Y.; Gao, J. AttranSUNet: An enhanced hybrid transformer architecture for ultrasound and histopathology image segmentation. *Comput. Biol. Med.* **2023**, *152*, 106365. [\[CrossRef\]](#)
28. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
29. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
30. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
31. Schmitz, R.; Madesta, F.; Nielsen, M.; Krause, J.; Steurer, S.; Werner, R.; Rösch, T. Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture. *Med. Image Anal.* **2021**, *70*, 101996. [\[CrossRef\]](#)
32. Srivastava, A.; Jha, D.; Chanda, S.; Pal, U.; Johansen, H.D.; Johansen, D.; Riegler, M.A.; Ali, S.; Halvorsen, P. MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2252–2263. [\[CrossRef\]](#)
33. Zhang, B.; Wang, Y.; Ding, C.; Deng, Z.; Li, L.; Qin, Z.; Ding, Z.; Bian, L.; Yang, C. Multi-scale feature pyramid fusion network for medical image segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 353–365. [\[CrossRef\]](#)
34. Yuan, Y.; Cheng, Y. Medical image segmentation with UNet-based multi-scale context fusion. *Sci. Rep.* **2024**, *14*, 15687. [\[CrossRef\]](#)
35. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In *Proceedings of the Artificial Intelligence and Statistics*; PMLR: New York, NY, USA, 2015; pp. 562–570.
36. Dou, Q.; Yu, L.; Chen, H.; Jin, Y.; Yang, X.; Qin, J.; Heng, P.-A. 3D deeply supervised network for automated segmentation of volumetric medical images. *Med. Image Anal.* **2017**, *41*, 40–54. [\[CrossRef\]](#)
37. Litjens, G.; Bandi, P.; Bejnordi, B.E.; Geessink, O.; Balkenhol, M.; Bult, P.; Halilovic, A.; Hermesen, M.; van de Loo, R.; Vogels, R.; et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience* **2018**, *7*, giy065. [\[CrossRef\]](#)
38. 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. Gland segmentation in colon histology images: The GlaS challenge contest. *Med. Image Anal.* **2017**, *35*, 489–502. [\[CrossRef\]](#)
39. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29.
40. Wang, H.; Cao, P.; Wang, J.; Zaiane, O.R. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2441–2449. [\[CrossRef\]](#)
41. Ruan, J.; Xiang, S. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. *arXiv* **2024**, arXiv:2402.02491.
42. Zhou, Y.; Li, L.; Lu, L.; Xu, M. nnWNet: Rethinking the Use of Transformers in Biomedical Image Segmentation and Calling for a Unified Evaluation Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 20852–20862.
43. Moshkin, A.; Fedorov, M.; Arlazarov, V.; Gribova, V.; Nazarenko, A.; Repin, D.; Klevtsova, O.; Romanov, A. Localization in Medical Imaging: A Unified AI Approach for Ovaries, Follicles, and Vertebral Arteries. *Algorithms* **2026**, *19*, 523. [\[CrossRef\]](#)
44. Gülgen, S. A Multimodal Deep Learning Framework for Personalized Event Recommendation with Hyperparameter Optimization via Genetic Algorithm. *J. Comput. Optim. Sci. Mach. Learn.* **2026**, *1*, 28–38. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.