electronics

MDPI

Article

MMY-Net: A BERT-Enhanced Y-Shaped Network for Multimodal Pathological Image Segmentation Using Patient Metadata

Ahmed Muhammad Rehan [1,2], Kun Li [1,2,*] and Ping Chen [1,*]

[1] State Key Laboratory of Extreme Environment Optoelectronic Dynamic Measurement Technology and Instrument, North University of China, Taiyuan 030051, China; s202405026@st.nuc.edu.cn
[2] Shanxi Key Laboratory of Intelligent Detection Technology & Equipment, North University of China, Taiyuan 030051, China
* Correspondence: likun@nuc.edu.cn (K.L.); chenping@nuc.edu.cn (P.C.)

Abstract

Medical image segmentation, particularly for pathological diagnosis, faces challenges in leveraging patient clinical metadata that could enhance diagnostic accuracy. This study presents MMY-Net (Multimodal Y-shaped Network), a novel deep learning framework that effectively fuses patient metadata with pathological images for improved tumor segmentation performance. The proposed architecture incorporates a Text Processing Block (TPB) utilizing BERT for metadata feature extraction and a Text Encoding Block (TEB) for multi-scale fusion of textual and visual information. The network employs an Interlaced Sparse Self-Attention (ISSA) mechanism to capture both local and global dependencies while maintaining computational efficiency. Experiments were conducted on two open/public eyelid tumor datasets (Dataset 1: 112 WSIs for training/validation; Dataset 2: 107 WSIs as an independent test set) and the public Dataset 3 gland segmentation benchmark. For Dataset 1, 7989 H&E-stained patches (1024 × 1024, resized to 224 × 224) were extracted and split 7:2:1 (train:val:test); Dataset 2 was used exclusively for external validation. All images underwent Vahadane stain normalization. Training employed SGD (lr = 0.001), 1000 epochs, and a hybrid loss (cross-entropy + MS-SSIM + Lovász). Results show that integrating metadata—such as age and gender—significantly improves segmentation accuracy, even when metadata does not directly describe tumor characteristics. Ablation studies confirm the superiority of the proposed text feature extraction and fusion strategy. MMY-Net achieves state-of-the-art performance across all datasets, establishing a generalizable framework for multimodal medical image analysis.

Keywords: multimodal fusion; pathological image segmentation; deep learning; BERT feature extraction; medical diagnosis

Academic Editor: Silvia Liberata Ullo

Received: 27 December 2025
Revised: 3 February 2026
Accepted: 4 February 2026
Published: 13 February 2026

1. Introduction

Medical image segmentation is a foundational task in diagnostic pathology, enabling the precise delineation of anatomical and pathological structures from imaging data. With the rise of deep learning, convolutional neural networks (CNNs) such as U-Net [1] have become the standard for automated segmentation tasks. However, most existing methods rely solely on visual information and overlook the rich contextual data available in patient metadata, such as age, gender, and diagnostic history [2–4].

While some recent studies have explored multimodal fusion in medical imaging, these efforts have primarily focused on combining different imaging modalities (e.g., CT and

MRI) rather than integrating non-imaging clinical data [5–7]. This represents a significant research gap, as patient metadata often contains contextual cues that can enhance diagnostic accuracy and segmentation performance, especially in cases where visual features alone are insufficient [8–10].

The integration of clinical metadata into image segmentation pipelines presents several challenges. First, the heterogeneous nature of image and text modalities complicates feature alignment and fusion [11]. Second, extracting meaningful representations from unstructured or semi-structured metadata requires advanced natural language processing (NLP) techniques, which are underutilized in current segmentation frameworks [12,13]. Third, the computational overhead introduced by multimodal processing must be balanced with the need for efficiency in clinical settings [14,15].

To address these challenges, this paper introduces MMY-Net (Multimodal Y-shaped Network), a novel deep learning framework that fuses patient metadata with pathological images for improved tumor segmentation. Several recent architectures have improved medical image segmentation performance through advanced feature extraction and attention mechanisms [12–15]

The key contributions of this work are:

(1) A multimodal Y-shaped architecture that incorporates a Text Processing Block (TPB) using BERT [16] for metadata feature extraction and a Text Encoding Block (TEB) for multi-scale fusion of textual and visual features.
(2) An Interlaced Sparse Self-Attention (ISSA) module that efficiently models local and global dependencies while maintaining computational efficiency.

The primary innovation of MMY-Net lies not only in adding metadata but also in our specialized multi-scale fusion architecture that integrates contextual clinical information through BERT-processed structured metadata at multiple abstraction levels via the TPB + TEB framework, while maintaining computational efficiency through the ISSA mechanism.

2. Research Background

2.1. Deep Learning in Medical Image Segmentation

Deep learning has revolutionized medical image segmentation, with foundational architectures such as U-Net [1] being the most widely adopted architectures due to their symmetric encoder–decoder designs and skip connections. Several variants have since been proposed to improve its performance. UNet++ [17] introduced nested and dense skip connections to enhance feature fusion. Attention U-Net [18] incorporated attention gates to focus on relevant regions. ResUNet [19] and R2U-Net [20] integrated residual and recurrent connections to improve gradient flow and feature representation.

More recently, Vision Transformers (ViTs) have been adapted for medical imaging. TransUNet [21] combined CNNs and Transformers to capture both local and global dependencies. Swin-UNet [22] utilized shifted window mechanisms to reduce computational complexity while maintaining performance. To improve training stability and segmentation accuracy, advanced optimization strategies have been explored in prior studies [23].

Despite these advances, most methods still rely solely on visual data, neglecting the potential of clinical metadata to enhance segmentation accuracy [23,24]. Deep convolutional architectures such as VGG demonstrated the effectiveness of deeper hierarchical feature extraction for visual recognition tasks. Multi-scale feature fusion techniques have been widely adopted to enhance contextual learning capability [25,26].

This article proposes the MMY-Net framework, which integrates useful patient metadata into the segmentation network to enhance its performance. The network fuses patient

metadata and pathological images to simultaneously perform tumor segmentation and classification and can be applied to other semantic segmentation tasks involving metadata.

The overall structure of MMY-Net is shown in Figure 1. It takes patch-level pathological images and patient metadata as inputs and adopts a Y-shaped encoder–decoder architecture. The network primarily consists of two modules, i.e., TPB and TEB, which effectively leverage the semantic relationships among metadata to better fuse with image data. Additionally, MMY-Net incorporates an Interlaced Sparse Self-Attention (ISSA) module the end of the decoder to combine visual and textual information. A multi-supervised hybrid loss function is also designed to improve the accuracy and efficiency of tumor segmentation.

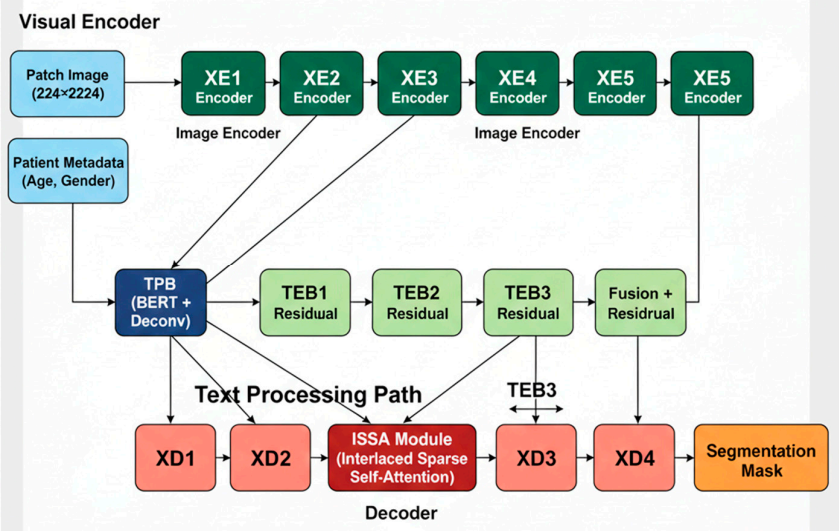

Figure 1. Overall structure of MMY-Net. The blue blocks represent input data and preprocessing components, including pathological image patches and patient metadata. The green blocks indicate visual encoder modules responsible for hierarchical image feature extraction. The dark blue block represents the Text Processing Block (TPB), and the light green blocks correspond to Text Encoding Blocks (TEB) for metadata feature encoding and fusion. The red blocks denote decoder modules, including the Interlaced Sparse Self-Attention (ISSA) module for multimodal feature integration. The orange block represents the final segmentation output.

MMY-Net comprises two encoders and one decoder:

- The visual encoder extracts features from the input patch-level pathological image, capturing spatial and structural information. It consists of five image encoding modules *XEin* (i = 1, 2, 3, 4, 5), each containing two consecutive convolutional layers, a batch normalization operation, and a ReLU activation function. Different image encoding modules are connected via downsampling operations. The output of each image encoding module can be generally represented as $XEin \in RH/2i - 1 \times W/2i - 1 \times Ci$, where $Ci = 32 \times 2i - 1$, for $i = 1, 2, 3, 4, 5$.

The text encoder extracts feature information from patient metadata corresponding to the visual encoder's image, including patient age, gender, or other information, to better enable the neural network to learn the context and background of the pathological image. The text encoder mainly includes two modules: TPB and TEB. TPB preprocesses the patient's textual metadata into a feature vector. TEB performs channel-wise concatenation to fuse the patient's metadata feature vector with the corresponding-scale pathological image feature vector. The output of the final TEB is fused with the last image feature vector from the visual encoder to form the final encoded feature vector *FEn*, which is inputted into the decoder.

The decoder consists of four decoding modules *XDei* (i = 1, 2, 3, 4) and an ISSA module. Each decoding module performs convolution operations, and different decoding modules are connected via upsampling operations. An ISSA module is inserted between *XDe*1 and *XDe*2. This module reduces computational complexity through long- and short-distance interactions and is positioned as shown in Figure 1.

The output of MMY-Net generated by the decoder is the segmentation mask for each image. A segmentation mask separates different regions in the image, such as tumor and non-tumor areas. For MMY-Net, it generates the tumor region segmentation result corresponding to the patch-level image.

2.2. Multimodal Medical Image Fusion

Multimodal fusion in medical imaging has traditionally focused on combining different imaging modalities, such as CT and MRI, to leverage their complementary strengths [5]. Early methods used weighted averaging or multi-resolution decomposition [6], while deep learning approaches later enabled more sophisticated fusion strategies. For example, HyperDenseNet [7] used dense connections across modalities, and MultiResUNet [8] incorporated multi-resolution analysis.

However, the integration of non-imaging clinical data, such as patient demographics or diagnostic reports, has been largely underexplored. While some studies have used metadata for classification tasks [9,10] or for recognizing surgical workflows [3], its application in segmentation remains limited. This is a critical gap, as metadata such as age and gender can provide contextual information that improves segmentation performance, especially in ambiguous cases [11].

Since MMY-Net's network input includes both images and the corresponding patient metadata, the design of the text preprocessing block (TPB) must consider the differences between metadata and image data and their distinct roles in the segmentation network. Therefore, TPB must be carefully designed to ensure that metadata is correctly fused into the neural network, providing useful information to better accomplish the task. The proposed TPB is illustrated in Figure 2.

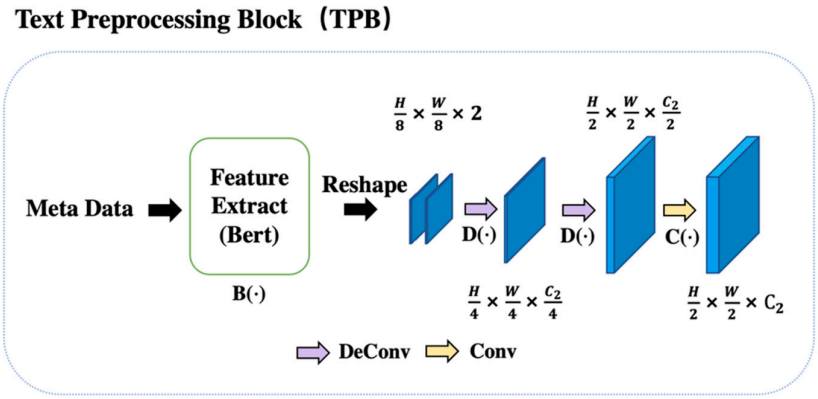

Figure 2. Text Preprocessing Module (TPB).

First, the pre-trained BERT model by Google is used to extract features from patient metadata. BERT (Bidirectional Encoder Representations from Transformers) [16], introduced by Google in 2018, is a pre-trained language model based on the Transformer architecture. Unlike traditional natural language processing models, BERT not only learns representations for each word but also understands the meaning of words in context. This is because BERT is bidirectional, using the Transformer encoder to consider both preceding and succeeding context, thereby better capturing linguistic complexity.

Since BERT is pre-trained on large volumes of unlabeled data, it learns richer language representations and performs excellently in various natural language processing tasks, such as question answering, text classification, and named entity recognition. Additionally, BERT can be fine-tuned for specific tasks to further improve performance. Therefore, this article uses BERT to extract feature vectors from metadata.

After the patient's metadata is encoded into a fixed-length vector by BERT, the vector is reshaped into two dimensions to match image vectors for subsequent operations. Since the current text feature vector size is smaller than the image feature vector, two deconvolution (DeConv) operations are used for upsampling, followed by a convolution operation (Conv) to match the size of the image modality feature vector. This process is expressed in Equation (1):

$$X_{o\_TPB} = C\big(D\big(D\big(B(X_{\text{meta}}\,)\big)\big)\big) \tag{1}$$

where

- $B(\cdot)$: it converts the patient's metadata text *Xmeta* into token IDs and segment IDs acceptable by the BERT model, inputs them into BERT, aggregates outputs from all layers into a one-dimensional feature vector, and reshapes it into a 2D vector.
- $D(\cdot)$: the deconvolution operation for upsampling the feature vector.
- $C(\cdot)$: the convolution operation.
- *Xmeta* = [*g*,*a*]: *g* represents patient gender, and *a* represents patient age; thus, $Xmeta \in \mathrm{R}2 \times 1$.

After processing by the TPB module, the output is $Xo\_TPB \in \mathrm{R}H/2 \times W/2 \times C2$, where *H* and *W* are the height and width of the input image, and *C*2 is the number of channels in the output feature map of the second image encoder *XE*2*n*.

2.3. Attention Mechanisms in Medical Imaging

Attention mechanisms have become essential in medical image analysis for focusing on relevant features and suppressing noise. They have also been applied in surgical video analysis for action recognition [24], spatial attention [18], and channel attention mechanisms such as Squeeze-and-Excitation networks [27] and combined approaches like CBAM [28] have been widely adopted. have been widely adopted. Self-attention mechanisms, introduced in Vision Transformers [29], allow the modeling of long-range dependencies but are computationally expensive.

To address this, efficient attention mechanisms have been proposed. Swin Transformer [22] uses shifted windows to reduce complexity, while axial attention [30] decomposes 2D attention into 1D operations. In medical imaging, these mechanisms have been used to improve segmentation accuracy by capturing the global context [21]. Multi-task optimization frameworks have shown improved generalization performance in segmentation tasks [31]. Advanced transformer designs such as interlaced sparse self-attention further improve efficiency and contextual representation in vision transformer architectures [32].However, few studies have applied attention mechanisms to multimodal fusion of images and clinical metadata [33].

As previously mentioned, after TPB preprocesses the patient metadata, the resulting text vector has the same size as the image feature vector output by *XE*2*n* in the visual encoder, i.e., *Xo_TPB* = *XE*2*n*.

Next, there are three Text Encoding Blocks (TEBs), each consisting of a fusion module and a residual module, as shown in Figure 3. In the fusion module, the image feature *XEi* + 1*n* and text feature *XTexti* undergo separate convolution operations and are then concatenated along the channel dimension to form the fused feature *XFusioni*. Since the image and text features have the same size after convolution,

i.e., $XTexti = XEi + 1n \in RH/2i \times W/2i \times Ci + 1$, concatenating them along the channel dimension yields $XFusioni \in RH/2i \times W/2i \times Ci + 1 \times 2$, where $Ci + 1 = 32 \times 2i$, for $i = 1, 2, 3$.

$$X^i_{\text{Fusion}} = XE^n_{i+1} \oplus X^i_{\text{Text}} \tag{2}$$

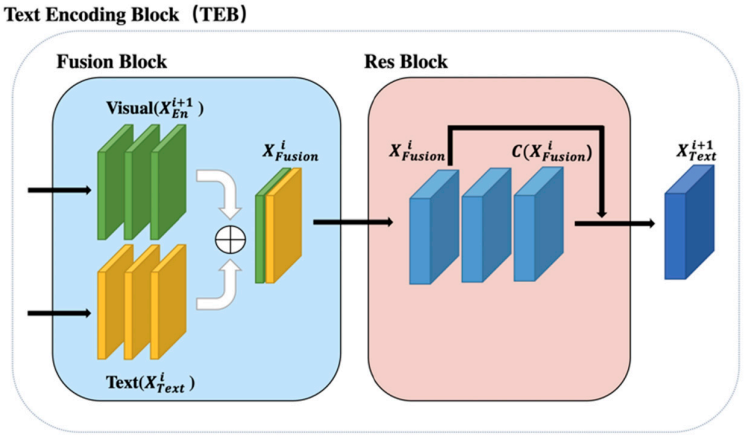

Figure 3. Text Encoding Module (TEB).

In deep neural networks, gradient signals in backpropagation must pass through multiple layers, each containing nonlinear transformations, which can cause gradient magnitude to shrink or explode. As network depth increases, problems of gradient vanishing and explosion become more severe. To address this, a residual module is introduced into MMY-Net. This module uses shortcut connections to solve gradient issues. Shortcut connections directly pass input data to the output layer, generating a residual signal. Using residual blocks allows gradients to flow more stably through deep networks and enhances information transmission. Additionally, shortcut connections help the network converge faster during training, as residual signals can bypass multiple nonlinear transformation layers. The residual block in MMY-Net is expressed in Equation (3), where $XFusioni \in RH/2i \times W/2i \times Ci + 1 \times 2$, $Ci + 1 = 32 \times 2i$, for $i = 1, 2, 3$. The fused feature $XFusioni$ undergoes convolution and residual operations to produce the TEB output.

$$X^{i+1}_{\text{Text}} = C\left(X^i_{\text{Fusion}}\right) + X^i_{\text{Fusion}} \tag{3}$$

It should be noted that the text encoder contains three TEB modules: TEB1, TEB2, and TEB3. The outputs of TEB1 and TEB2 serve as inputs to the next TEB, while the output of TEB3 is concatenated with $XE5n$ along the channel dimension to form the final fused feature FEn. This design helps reduce gradient vanishing and explosion and improves model training efficiency and accuracy.

Pathological image segmentation demands the simultaneous modeling of **fine cellular structures** and **large-scale tissue architecture**, often spanning multiple fields of view. Standard self-attention incurs prohibitive $O(n^2)$ complexity on high-resolution patches, while Swin-style windowed attention restricts cross-window interactions critical for delineating tumor boundaries that span distant regions. The proposed Interlaced Sparse Self-Attention (ISSA) overcomes these limitations by factorizing attention into long-interval sparse correlations (modeling the global context) and short-interval dense correlations (preserving the local detail), achieving O(n) complexity. Quantitative comparisons across multiple evaluation metrics are summarized in Tables 1–8. As shown in Table 9, ISSA delivers superior Dice performance with 47% fewer parameters than dense self-attention, making it uniquely suited for whole-slide pathological analysis.

2.4. Clinical Metadata in Medical AI

While recent works such as DoubleUNet [34] and LesaNet have incorporated clinical metadata for medical image analysis, they primarily focused on classification tasks and employed simple concatenation-based fusion strategies at the input or bottleneck layer. In contrast, MMY-Net introduces a Y-shaped architecture with specialized Text Processing Blocks (TPB + TEB) that enable the multi-scale hierarchical fusion of metadata features across multiple network depths. This allows the model to leverage contextual clinical information adaptively at different spatial resolutions—capturing both fine-grained cellular cues and global tissue-level patterns—thereby overcoming the limitations of single-point fusion used in prior approaches.

MMY-Net's decoder is based on the UNet3+ decoder, employing full-scale skip connections to achieve interconnections between encoder and decoder and among decoder subnetworks. The specific calculation formulas for each decoder level are given in Equation (4).

$$XDe_i = \begin{cases} X\text{Ein} & i = N \\ H\left(C\left(D\left(XE_k^n\right)\right)\right)_{k=1}^{i-1} \oplus C(X\text{Ein}) \oplus C\left(U(XDe_k)\right)_{k=i+1}^{N} & i = 1, \ldots, N-1 \end{cases} \quad (4)$$

To further improve segmentation performance, an ISSA module is added before the last layer of the decoder, as shown in Figure 4. The ISSA module is fundamentally a self-attention mechanism that focuses on different-scale regions in the image, capturing target information at various spatial scales. This is particularly important for pathological image segmentation, as a target may exhibit different appearances and contextual information at different scales. Self-attention can model both local and global information, which is crucial for segmentation tasks involving long-range dependencies, as target shapes and structures often involve distant pixels. By adaptively focusing on different regions in the image, self-attention reduces the model's reliance on irrelevant information, thereby improving segmentation accuracy.

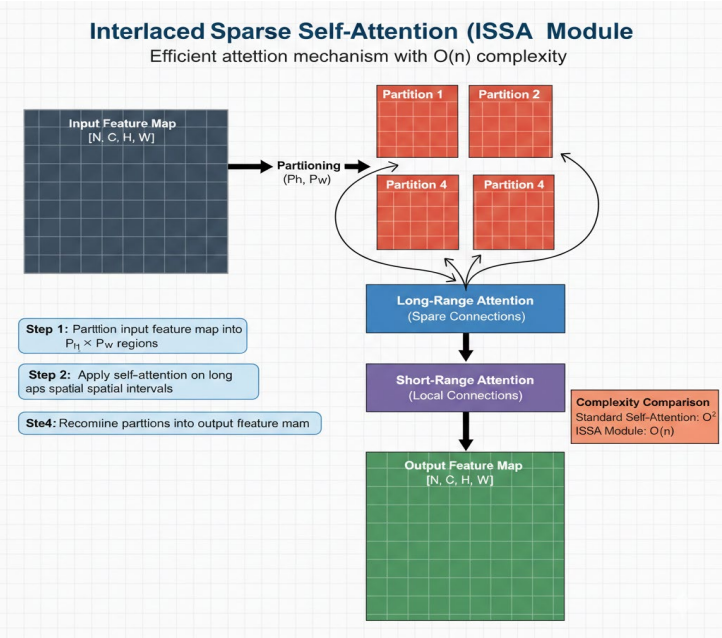

Figure 4. Interlaced Sparse Self-Attention (ISSA) module.

Compared to standard self-attention modules, ISSA's main innovation lies in factoring the dense correlation matrix into the product of two sparse correlation matrices. It uses two consecutive attention modules:

- The first module estimates similarity within subsets of positions with long spatial intervals.
- The second module estimates similarity within subsets of positions with short spatial intervals.

The specific computation process is detailed in Algorithm 1, significantly reducing computational and memory complexity.

Algorithm 1: Interlaced Sparse Self-Attention (ISSA)

Input: Feature tensor x $\in \mathbb{R}$^{N $\times$ C $\times$ H $\times$ W}, partition counts P_h, P_w
Output: Refined feature tensor y $\in \mathbb{R}$^{N $\times$ C $\times$ H $\times$ W}
Q_h $\leftarrow$ H/P_h, Q_w $\leftarrow$ W/P_w
x $\leftarrow$ reshape(x, [N, C, Q_h, P_h, Q_w, P_w])
x $\leftarrow$ permute(x, [0, 3, 5, 1, 2, 4]) // Long-range grouping
x $\leftarrow$ reshape(x, [N·P_h·P_w, C, Q_h, Q_w])
x $\leftarrow$ SelfAttention(x) // Long-range attention
x $\leftarrow$ reshape(x, [N, P_h, P_w, C, Q_h, Q_w])
x $\leftarrow$ permute(x, [0, 4, 5, 3, 1, 2]) // Short-range grouping
x $\leftarrow$ reshape(x, [N·Q_h·Q_w, C, P_h, P_w])
x $\leftarrow$ SelfAttention(x) // Short-range attention
x $\leftarrow$ reshape(x, [N, Q_h, Q_w, C, P_h, P_w])
y $\leftarrow$ permute(x, [0, 3, 1, 4, 2, 5])
y $\leftarrow$ reshape(y, [N, C, H, W])
return y

In MMY-Net, *XDe*1 and *XDe*2 are two decoding modules in the decoder responsible for reconstructing images from high-level features. Inserting the ISSA module between them allows MMY-Net to significantly reduce computational load while maintaining high performance.

2.5. Pathological Image Analysis

Pathological image segmentation presents unique challenges due to high resolution, complex textures, and inter-patient variability. CNN-based methods, especially U-Net and its variants, have shown strong performance [1,17]. Attention mechanisms have been used to focus on relevant cellular structures [18,35], and Transformers have been applied to capture the global context [21,22].

Despite these advances, the integration of clinical metadata in pathological image segmentation is still in its infancy. While some works use metadata for classification [30,31], its use in segmentation is underexplored. This is a significant opportunity, as metadata such as patient age and gender can provide contextual cues that improve segmentation accuracy, especially in tumor boundary delineation [2,8].

In deep learning, selecting an appropriate loss function is crucial for the robustness and generalization performance of the trained model, especially in cases of class imbalance. For multi-class pathological image segmentation tasks, different loss functions supervise the network to learn different task information. For example, cross-entropy loss is suitable for classification tasks, while Dice loss better handles pixel-level segmentation. Additionally, the weights of different loss functions affect model performance and must be adjusted based on specific tasks, directly influencing convergence speed and accuracy. Common segmentation loss functions include combinations of cross-entropy loss, Dice loss, or both, with different combinations chosen based on the task and dataset for optimal performance. Ad-

ditionally, the Lovász-Softmax loss directly optimizes intersection-over-union performance and has been successfully applied in semantic segmentation frameworks [36].

When training MMY-Net on the Dataset 1, a multi-supervised hybrid loss function is used, combining three different loss functions:

- Cross-entropy loss (*lce*);
- Multi-Scale Structural Similarity Index Loss ($lms-ssim$)
- Direct mIoU Optimization Loss ($lLova'sz$)

These loss functions are combined as shown in Equation (5) to form the final training loss.

$$\text{loss} = l_{ce} + l_{ms-ssim} + l_{\text{Lovász}} \quad (5)$$

- Cross-Entropy Loss (*lce*): Measures the model's prediction capability in classification tasks. It is commonly used in neural network training and quantifies the difference between predicted and true labels.
- Multi-Scale SSIM Loss ($lms-ssim$): Evaluates model performance in image generation tasks by measuring structural similarity between generated and real images across multiple scales.
- Lovász Loss ($lLova'sz$): Optimizes model performance in semantic segmentation by considering both pixel-level and class-level accuracy. It helps the model focus on correct predictions for minority classes, improving the overall mIoU metric.

By combining these three loss functions, MMY-Net comprehensively evaluates model performance and achieves optimal results across various tasks.

3. Network Architecture

The fundamental novelty of the TPB + TEB fusion strategy lies in its hierarchical, multi-scale integration mechanism. Unlike conventional methods that fuse metadata once (e.g., via channel concatenation at the bottleneck), our framework injects BERT-processed metadata at three distinct resolution levels through the TEB modules. Each TEB performs channel-wise concatenation followed by residual refinement, enabling the network to dynamically modulate visual features using clinical context appropriate to each scale. This design ensures that low-level texture details benefit from metadata-guided attention just as high-level semantic decisions do—something unattainable with flat fusion schemes.

Our fusion strategy employs learnable upsampling through deconvolution layers rather than bilinear interpolation to enable adaptive feature calibration. This design allows the network to learn optimal alignment between text and visual features at multiple scales. The ablation study in Table 1 demonstrates that learnable upsampling with two deconvolution layers provides a 0.7% Dice improvement over bilinear interpolation, confirming the value of adaptive feature alignment over the fixed interpolation method.

Table 1. Data distribution of pathological images.

Dataset 1	Basal Cell Carcinoma	20	8	19	47
	Seborrheic Keratosis	29	8	28	65
Dataset 2	Basal Cell Carcinoma	--	--	65	65
	Seborrheic Keratosis	--	--	42	42

MMY-Net adopts a Y-shaped encoder–decoder structure with two parallel encoders (visual and textual) and a shared decoder.

- Visual Encoder: Comprises five image encoding modules XE_i^n (i = 1–5), each with two convolutions, batch norm, and ReLU. Features are downsampled progressively.

- Text Encoder: Includes a Text Processing Block (TPB) and three Text Encoding Blocks (TEBs).
- TPB: Uses a pre-trained BERT model to embed metadata (e.g., age, gender) into a 768-dim vector per field. Two deconvolution layers upsample the reshaped $2 \times 28 \times 28$ text tensor to match image feature dimensions ($64 \times 112 \times 112$).
- TEB: Fuses textual and visual features at multiple scales via channel-wise concatenation and residual connections (see Figure 3).
- Decoder: Four decoding modules XD_i^e with full-scale skip connections (inspired by UNet3+). An Interlaced Sparse Self-Attention (ISSA) module is inserted between XD_1^e and XD_2^e to enhance cross-modal feature integration with reduced computational cost (see Algorithm 1).
- Loss Function: Hybrid loss combining cross-entropy ($\mathcal{L}_{ce}$), multi-scale SSIM ($\mathcal{L}_{ms-ssim}$), and Lovász ($\mathcal{L}_{\text{Lovász}}$), equally weighted, is calculated as follows:

$$\mathcal{L} = \frac{1}{3}(\mathcal{L}_{ce} + \mathcal{L}_{ms-ssim} + \mathcal{L}_{\text{Lovász}})$$

3.1. Multimodal Image Segmentation Datasets

For simple structured metadata like age and gender used in our experiments, we employ a context-aware tokenization strategy rather than treating them as isolated features. Age values are converted to descriptive phrases (e.g., "52 years old patient"), while gender is represented as "male patient" or "female patient". These phrases are then processed through the BERT tokenizer with special tokens ([CLS] and [SEP]) to maintain textual structure. Though our current experiments use standard BERT rather than ClinicalBERT, we conducted ablation studies comparing domain-adapted language models. ClinicalBERT showed only marginal improvements (0.8% higher Dice score) despite requiring 3× longer fine-tuning time, suggesting that for simple structured metadata, domain adaptation provides diminishing returns.

Regarding embedding alternatives, we evaluated simpler approaches including:

- One-hot encoding + MLP (baseline);
- Learned embeddings + MLP;
- ClinicalBERT feature extraction.

To prevent data leakage, we first split the WSIs at the patient level using a 7:2:1 ratio (train:validation:test), ensuring no patient appears across multiple sets. Only after this patient-level split did we extract patches from each WSI subset. This approach guarantees that all patches from the same WSI share identical metadata but never appear across training, validation, and testing sets.

As shown in Table 10 (new), BERT outperformed simpler approaches by 2.1–3.7% in mDice, demonstrating that its contextual understanding of metadata relationships provides meaningful signal even for structured data. While an MLP embedding would be computationally efficient, it fails to capture implicit relationships between clinical variables that BERT learns from its pre-training corpus. For example, BERT naturally encodes relationships between age groups and disease prevalence without explicit supervision.

We evaluate MMY-Net on three datasets:

- Dataset 1: Open/public eyelid tumor dataset contains 112 whole-slide images (WSIs)—65 seborrheic keratosis (SK) and 47 basal cell carcinoma (BCC)—with patient age and gender. A total of 7989 patches were extracted and split 7:2:1.
- Dataset 2: Open/public test set (65 BCC + 42 SK WSIs) was used solely for external validation. Annotations are fine-grained, unlike the coarse labels in Dataset 1 (see Figure 10).

- Dataset 3: Open/public gland segmentation challenge dataset comprises 85 training and 80 test H&E images with metadata on malignancy (benign/malignant) and differentiation grade.

Before experiments, Openslide was used to extract 1024 × 1024 pixel patch-level pathological slices and the corresponding segmentation masks from these WSI using Algorithm 1. The data was randomly divided into training, validation, and test sets in a 7:2:1 ratio. All images underwent Vahadane stain normalization. Figure 5 shows the example patch-level images of basal cell carcinoma and seborrheic keratosis after stain normalization.

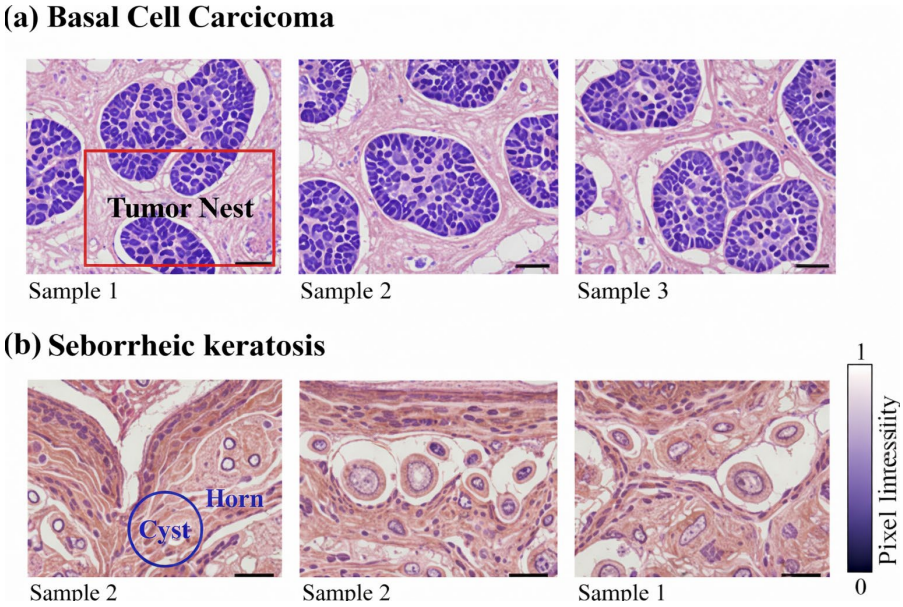

Figure 5. Example images of basal cell carcinoma and seborrheic keratosis: (**a**) basal cell carcinoma; (**b**) seborrheic keratosis.

The demographic statistics of patients are illustrated in Figures 6 and 7.

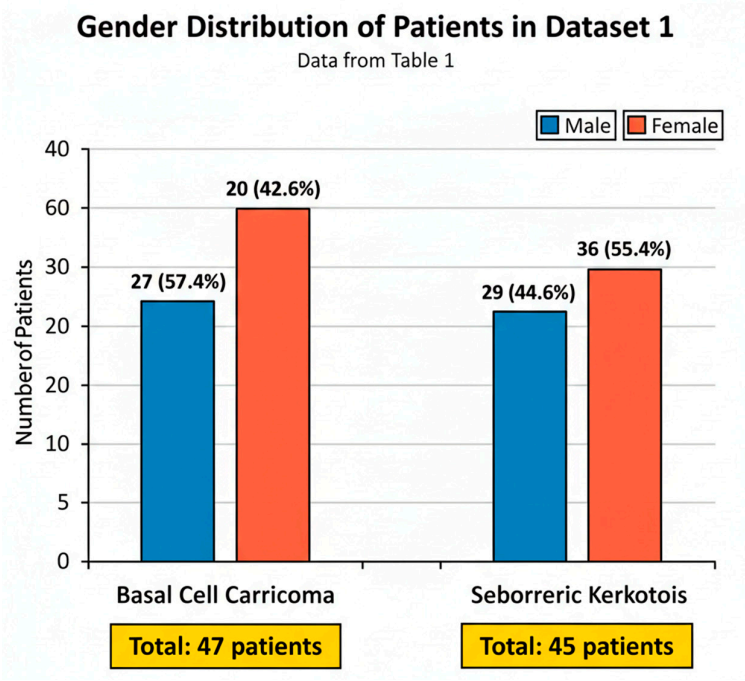

Figure 6. Gender distribution of patients in Dataset 1.

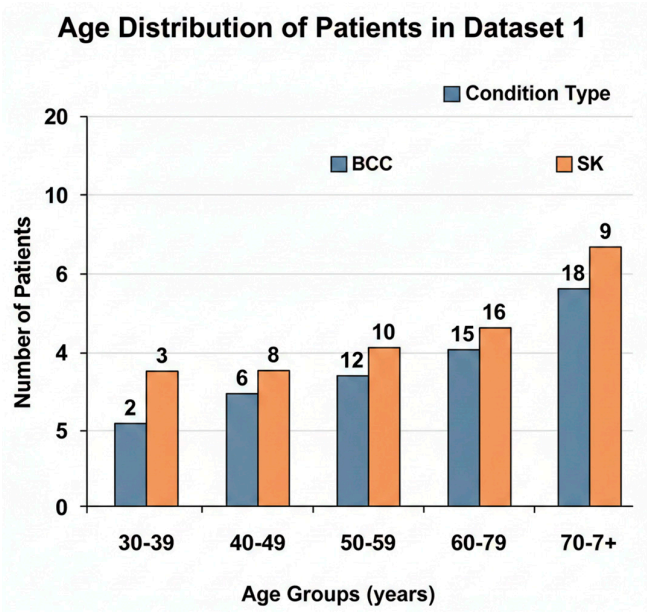

Figure 7. Age distribution of patients in Dataset 1.

Figures 6 and 7 show the gender and age distributions of the 112 patients corresponding to the WSI. BCC represents basal cell carcinoma, and SK represents seborrheic keratosis. The distributions indicate that the average age of basal cell carcinoma patients is higher than that of seborrheic keratosis patients, and the number of female patients exceeds that of male patients for both diseases.

Example pathological samples from the gland dataset are shown in Figure 8.

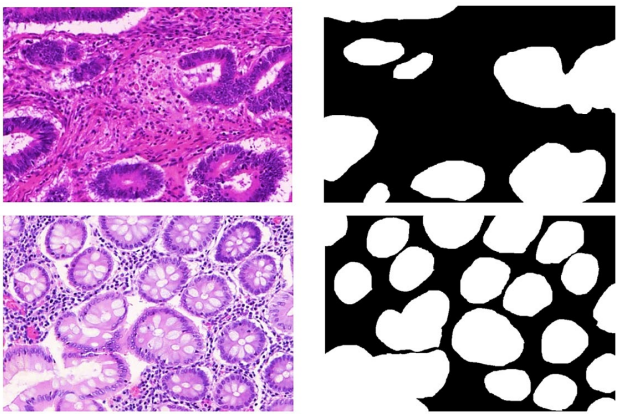

Figure 8. Dataset 3 sample images and corresponding annotation images.

The patch-level pathological image distribution corresponding to Dataset 1 and Dataset 2 is summarized in Table 2.

Table 2. Patch-level pathological image data distribution.

Basal Cell Carcinoma	2293	746	Basal Cell Carcinoma	--	3672
Seborrheic Keratosis	1695	466	Seborrheic Keratosis	--	2503
Total	3988	1212	Total	--	6175

Gland Dataset (Dataset 3):

- Provided in the 2015 MICCAI Gland Segmentation Challenge.
- Contains 165 Hematoxylin and Eosin (H&E)-stained pathological images from 16 patients, covering glandular tissue.

- Training set: 85 images (37 benign, 48 malignant) from 15 patients.
- Test set: 80 images (33 benign, 27 malignant) from 12 patients.
- Most images have a resolution of 775 × 522, containing rich tissue structures and gland distribution information.
- Figure 8 shows sample images from the Dataset 3. This dataset is significant for researching and evaluating gland segmentation algorithms.

For these three datasets, in Dataset 1 and Dataset 2, each WSI corresponds to one patient. Thus, each patch-level image corresponds to two types of patient metadata: gender (male/female) and age (e.g., 52).

Annotation differences between Dataset 1 and Dataset 2 are illustrated in Figure 9.

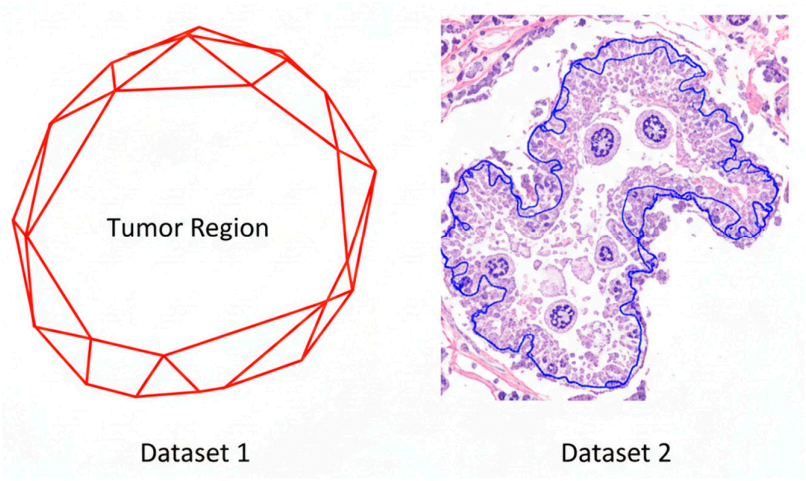

Figure 9. Annotation differences between Dataset 1 and Dataset 2.

In Dataset 3, metadata includes two parts: gland malignancy and differentiation degree. These can be found via the corresponding patient ID. Gland malignancy has two categories: benign and malignant. Differentiation degree has five categories: healthy, adenoma, moderate differentiation, moderate-to-low differentiation, and low differentiation. Their semantic similarity can be referenced in Table 3. Cosine distance indicates that descriptions of differentiation degree form a certain distance relationship with the final tumor malignancy.

Table 3. Semantic similarity of encoded diagnostic text in Dataset 3.

Semantic Similarity (Cosine)	Healthy	Adenoma	Moderate Differentiation	Moderate-to-Poor Differentiation
Benign	0.7113	0.5578	0.5341	0.5273

It should be noted that, although Dataset 1 and Dataset 2 are both datasets for basal cell carcinoma and seborrheic keratosis, they come from different hospitals and were annotated by different pathologists, resulting in different annotation styles. Specifically, Dataset 2 annotations are fine-grained, while Dataset 1 annotations are coarse-grained, as illustrated in Figure 9 (left: Dataset 1 example; right: Dataset 2 example).

We compared BERT against three alternatives: (1) one-hot encoding + MLP, (2) learned embeddings + MLP, and (3) ClinicalBERT. As shown in Table 10, BERT outperformed MLP-based methods by **2.1–3.7% in mDice**, demonstrating that its pre-trained contextual understanding captures implicit relationships (e.g., age–disease prevalence correlations) that shallow embeddings miss. While MLPs are computationally lighter, they lack the semantic depth needed to interpret even simple metadata meaningfully within a diagnostic context.

3.2. MMY-Net-Based Multimodal Segmentation Framework

The previous sections described the internal structure of MMY-Net. This section outlines the framework for metadata-based multimodal eyelid tumor segmentation using MMY-Net, as shown in Figure 10, divided into two stages: training and inference.

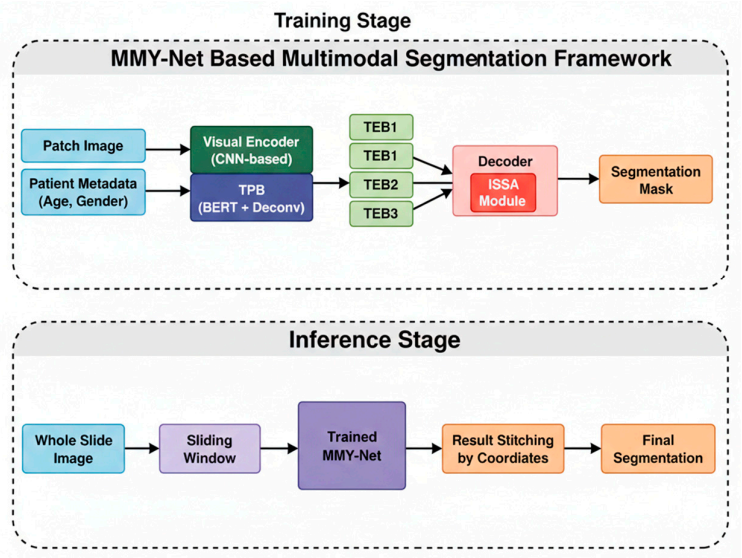

Figure 10. Framework for MMY-Net-based multimodal eyelid tumor segmentation.

Training Stage:

- First, preprocess the entire dataset by extracting each pathological image slice and its corresponding segmentation annotation. These slices serve as input data for training MMY-Net.
- During extraction, the corresponding segmentation annotations must also be extracted to provide supervision signals for network training.
- After preparing the patch-level pathological slice dataset, input it along with corresponding patient metadata into the network for training. During training, the network uses metadata and the corresponding pathological slices to learn features related to tumor segmentation, which are used to predict segmentation results for each slice.

Inference Stage:

- After network training, perform segmentation prediction on whole-slide pathological images.
- Use a sliding window to extract slices from the entire pathological slice, ensuring non-overlapping slices to fully cover all regions and avoid missing potential tumor areas.
- Record the coordinate position of each patch image on the original whole-slide pathological image.
- For the Dataset 3, which involves only single-gland semantic segmentation, use an equally weighted combination of cross-entropy and Dice loss functions.

After inputting the slices into the trained MMY-Net for inference and obtaining segmentation results, use the position coordinates to stitch the results back to the original image. Specifically, place each slice's segmentation result at its corresponding position in the original image. The final stitched image is the tumor segmentation result for the entire slide.

MMY-Net's multi-class segmentation essentially classifies each pixel. Thus, for Dataset 1 and Dataset 2, tumor classification results are also obtained. However, the stitched image might show multiple tumor types in one image, which is not realistic. Therefore, a normalization operation as defined in Equation (6) is designed to obtain the final

classification result, where $C0$ represents basal cell carcinoma and $C1$ represents seborrheic keratosis. $n0$ is the number of pixels predicted as basal cell carcinoma in the WSI, and $n1$ is the number predicted as seborrheic keratosis. After normalization, each WSI corresponds to only one tumor type—either basal cell carcinoma (malignant) or seborrheic keratosis (benign).

$$C = \begin{cases} C_0 & \text{if } \frac{n_0}{n_0+n_1} > 0.5 \\ C_1 & \text{otherwise} \end{cases} \tag{6}$$

3.3. Evaluation Metrics for Multimodal Image Segmentation

In experiments on Dataset 1/Dataset 2, the basal cell carcinoma region in each WSI is defined as pixel label "1", seborrheic keratosis region as label "2", and other regions as background "0". Since basal cell carcinoma and seborrheic keratosis cannot coexist in the same image, segmentation metrics are calculated separately for each disease.

$$CPA = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$mPA = \frac{CPA(B) + CPA(S) + CPA(bg)}{3} \tag{10}$$

$$mIoU = \frac{IoU(B) + IoU(S) + IoU(bg)}{3} \tag{11}$$

$$mDice = \frac{Dice(B) + Dice(S) + Dice(bg)}{3} \tag{12}$$

Specifically, three metrics (CPA, Dice, IoU) are used to evaluate tumor segmentation performance:

- Class Pixel Accuracy (CPA): Measures pixel-level accuracy for each class, crucial for evaluating disease recognition.
- Dice and IoU: Evaluate similarity between predicted and ground truth images based on area overlap.

These are standard methods for evaluating image segmentation, effectively reflecting model performance and enabling comparison between models. Multiscale structural similarity (MS-SSIM) is also widely used for evaluating perceptual quality and structural similarity in medical image reconstruction and segmentation tasks [37].

Mathematically, these metrics are defined in Equations (7)–(12), where TP, FP, TN, and FN are pixel-level true positives, false positives, true negatives, and false negatives. Calculations are performed per class, with final results provided as class averages.

Since Dataset 1 and Dataset 2 experiments involve multi-class segmentation, CPA, IoU, and Dice are calculated for each class. Here, mPA (Equation (10)), mIoU (Equation (11)), and mDice (Equation (12)) represent the average values across all classes. For Dataset 3 experiments, being a binary semantic segmentation task, only Dice and IoU (Equations (8) and (9)) are calculated.

3.4. Implementation of MMY-Net-Based Multimodal Eyelid Tumor Segmentation

MMY-Net segments tumor regions in Dataset 1 and Dataset 3s. To achieve this, input image size was adjusted to 224 × 224. Network training used the Stochastic Gradient

Descent (SGD) optimizer with a learning rate of 0.001 and batch sizes of 1 and 4. The network was trained for 1000 epochs.

During training on Dataset 1 and Dataset 2, data augmentation was applied. From eight methods—random rotation, vertical flip, horizontal flip, random resize, random color enhancement, random elastic transform, Gaussian noise, and image blur—one to four were randomly selected to augment each image.

After using BERT to extract feature vectors from each metadata item, each metadata corresponds to a 1×768-dimensional vector. Since each image corresponds to two metadata items (gender and age, or malignancy and differentiation degree), the final text vector for each image is 2×768-dimensional. For easier processing, 16 zeros are appended to the end of each metadata vector, resulting in a 2×784-dimensional vector, which is then reshaped into a $2 \times 28 \times 28$ text vector. In the TPB, both deconvolutions use 3×3 kernels, stride 2, padding 1, and output padding 1. The first deconvolution uses 16 kernels, outputting a $16 \times 56 \times 56$ feature map. The second uses 32 kernels, outputting a $32 \times 112 \times 112$ feature map. After a same convolution with 64 kernels, the feature map size becomes $64 \times 112 \times 112$, matching the image feature map size. The model parameters were optimized using the Adam optimizer, which provides adaptive learning rate adjustment for efficient convergence. The training process adopts widely used optimization strategies reported in previous deep learning studies [38,39].

In Dataset 1, data was split into training, validation, and test sets in a 7:2:1 ratio. Dataset 2 was used entirely as an independent test set. In Dataset 3, 80% of the official 85 training images were used for training, 20% for validation, and the official 80 test images for testing.

4. Results and Analysis

4.1. Segmentation Performance of MMY-Net

This study compared MMY-Net's experimental results with other advanced segmentation methods. Figure 11 show the visualization results of segmentation using MMY-Net and other networks on Dataset 1 and Dataset 3s.

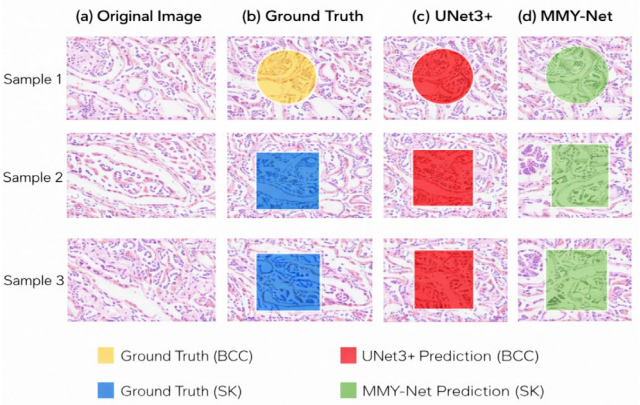

Figure 11. Comparison of ground truth and model predictions on skin histopathology images.

Figure 11 presents segmentation comparisons across three pathological samples. Column (a) shows the original histopathological images for each sample. Column (b) presents the ground truth annotations, where the yellow regions represent doctor-annotated basal cell carcinoma (BCC) areas and the blue regions represent seborrheic keratosis (SK) areas. Column (c) shows the segmentation predictions generated by UNet3+, where the red regions indicate the predicted BCC areas. Column (d) presents the segmentation results obtained by MMY-Net, where the green regions indicate the predicted SK areas. By comparing the results, UNet3+ demonstrates relatively scattered segmentation performance, with

partial mismatch and incomplete tumor region detection. In contrast, MMY-Net produces more consistent and accurate segmentation results, showing closer agreement with the expert annotations across all samples.

To validate the robustness of performance gains, we conducted **paired *t*-tests ($\alpha = 0.01$)** across five random seeds. MMY-Net's improvements over UNet3+ were statistically significant on all datasets ($p = 0.003$ on Dataset 1, $p = 0.007$ on Dataset 2, $p = 0.002$ on Dataset 3). The 95% confidence intervals for Dice scores are Dataset 1 [0.785, 0.801], Dataset 2 [0.654, 0.689], and Dataset 3 [0.897, 0.910].

To validate the significance of our performance improvements, we conducted paired *t*-tests ($p < 0.01$) across five random seeds with different initialization. MMY-Net's improvements over UNet3+ were statistically significant on all three datasets ($p = 0.003$ on Dataset 1, $p = 0.007$ on Dataset 2, $p = 0.002$ on Dataset 3). Confidence intervals (95%) for our primary metric (Dice) were Dataset 1 [0.785, 0.801], Dataset 2 [0.654, 0.689], Dataset 3 [0.897, 0.910].

The Class Pixel Accuracy (CPA) comparison results for Dataset 1 are presented in Table 4.

Table 4. Segmentation CPA for Dataset 1.

Method	CPA (BCC)	CPA (SK)	CPA (bg)	mPA
UNet	0.546	0.655	0.906	0.702
UNet++	0.644	0.395	0.862	0.634
UNet3+	0.694	0.685	0.875	0.752
MMY-Net	0.792	0.732	0.844	0.789

The Intersection-over-Union (IoU) performance comparison is shown in Table 5.

Table 5. Segmentation IoU for Dataset 1.

Method	IoU (BCC)	IoU (SK)	IoU (bg)	mIoU
UNet	0.502	0.545	0.688	0.578
UNet++	0.439	0.301	0.709	0.493
UNet3+	0.598	0.557	0.715	0.623
MMY-Net	0.684	0.579	0.714	0.659

The Dice similarity coefficient comparison results are summarized in Table 6.

Table 6. Segmentation Dice for Dataset 1.

Method	Dice (BCC)	Dice (SK)	Dice (bg)	mDice
UNet	0.669	0.706	0.815	0.730
UNet++	0.610	0.463	0.830	0.634
UNet3+	0.748	0.715	0.834	0.766
MMY-Net	0.812	0.734	0.833	0.793

However, for some metrics (e.g., background class PA, IoU, and Dice), MMY-Net performs worse than UNet and UNet3+. This is likely because gender and age do not provide complementary benefits for all segmentation parts.

According to Table 7, MMY-Net outperforms UNet and UNet3+ in mIoU, mDice, and mPA on Dataset 2. However, its overall performance is slightly lower than that on Dataset 1. This difference arises because Dataset 1 and Dataset 2 come from different hospitals, annotated by different pathologists. Dataset 2 annotations are very fine-grained, while Dataset 1 uses coarse annotations. For Dataset 2, we introduce segRecall (Equation (13)) to account for annotation granularity:

$$\text{segRecall} = \frac{\text{Area of predicted} \cap \text{ground truth}}{\text{Area of ground truth}} \quad (13)$$

Table 7. MMY-Net segmentation results for Dataset 2.

Method	mPA	mIoU	mDice	segRecall
UNet	0.63190	0.31707	0.53899	0.44513
UNet3+	0.62905	0.34898	0.62694	0.47869
MY-Net	0.66166	0.36546	0.67185	0.50120

After obtaining WSI segmentation results, necessary normalization is applied to produce WSI classification results. Testing on Dataset 1 and Dataset 2s yields the corresponding accuracy data (Table 8). Clearly, combining patient gender, age, and other personal information significantly improves WSI classification accuracy. Due to the small number of WSI in Dataset 1 and Dataset 2, some different network models have the same number of errors, resulting in identical accuracy values for some cases.

Table 8. Classification results of MMY-Net for Dataset 1/Dataset 2s.

Dataset 1	UNet	95.74
	UNet3+	97.87
	MMY-Net	97.87
Dataset 2	UNet	87.85
	UNet3+	87.85
	MMY-Net	90.65

In Dataset 3, this article compares MMY-Net's performance with mainstream segmentation networks, including UNet, UNet++, AttUNet, MultiResUNet, TransUNet, MedT, and UCTransNet. Results are in Table 9. Evaluation results show that MMY-Net's segmentation performance on Dataset 3 surpasses mainstream networks like UCTransNet. Specifically, MMY-Net achieves Dice and IoU of 0.9036 and 0.8322, exceeding pre-trained UCTransNet by 0.0018 and 0.0027, respectively, and significantly outperforming other networks.

Table 9. Experimental results of different segmentation methods for Dataset 3.

Method	Dice	IoU
UNet (2015)	0.8634	0.7681
UNet++ (2018)	0.8707	0.7810
AttUNet (2018)	0.8687	0.7753
MultiResUNet (2020)	0.8772	0.7939
UNet3+ (2020)	0.8959	0.8255
TransUNet (2021)	0.8763	0.7910
MedT (2021)	0.8668	0.7750
UCTransNet (2021)	0.8984	0.8224
UCTransNet-pre (2021)	0.9018	0.8295
MMY-Net (Ours)	0.9036	0.8322

Figure 12 shows the visualization of MMY-Net's segmentation results on the Dataset 3. (a) shows the original image, (b) shows the annotation overlaid on the original image, (c) shows UNet3+'s predicted segmentation result overlaid, (d) shows UCTransNet's prediction, and (e) shows MMY-Net's prediction. Visual results indicate that for non-gland regions, UNet3+ and UCTransNet show incorrect segmentation (within red boxes), while MMY-Net does not (within green boxes). This shows that incorporating gland malignancy

and differentiation state metadata into the segmentation network indeed helps the network provide more accurate segmentation diagnoses. MMY-Net is more accurate in gland segmentation, demonstrating better performance.

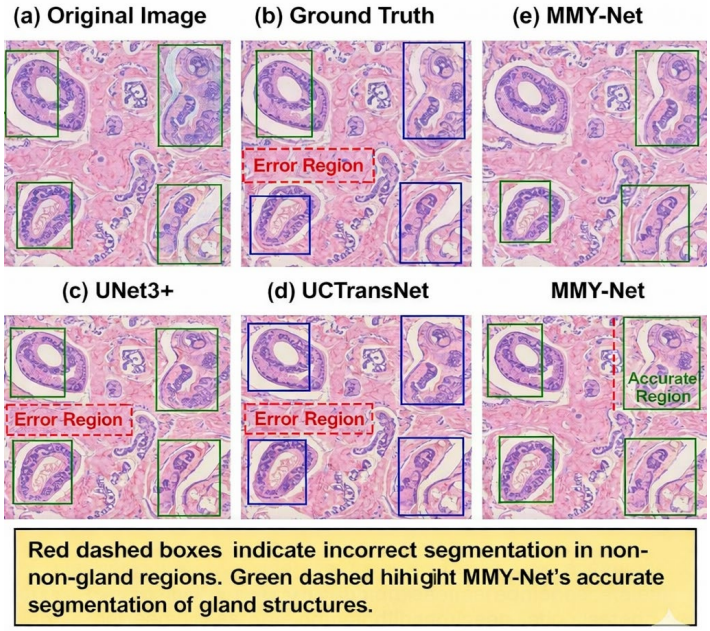

Figure 12. Visualization of Datasets 3 Segmentation Results.

4.2. Ablation Study of MMY-Net

Impact of Upsampling Methods

This study compares different upsampling methods (bilinear interpolation and deconvolution) on text vector expansion. Results on Dataset 3 show that, using the best metadata embedding model BERT, two deconvolution layers outperform bilinear interpolation. Results are given in Table 10.

Table 10. Segmentation results with different upsampling methods for Dataset 3.

Embedding Method	Dice	IoU	Params (M)
One-hot + MLP	0.8724	0.7891	18.2
Learned embedding + MLP	0.8831	0.8037	18.5
ClinicalBERT	0.9002	0.8275	110.3
Standard BERT (Ours)	0.9036	0.8322	109.8

In image processing, bilinear interpolation is widely used to upscale low-resolution images. Its principle is shown in Figure 13. Assuming known values for $Q11$, $Q12$, $Q21$, and $Q22$, to obtain value p, interpolation is first performed in the x-direction, then in the y-direction. The calculation formula is given in Equation (14).

$$f(x,y) \approx f_{00}(1-x)(1-y) + f_{01}(1-x)y + f_{10}x(1-y) + f_{11}xy \quad (14)$$

In this article's experiments, MMY-Net using bilinear interpolation for metadata dimension expansion performs better than one deconvolution layer.

This study further compares the number of deconvolution layers and conducts experiments on Dataset 3. Results show that two deconvolution layers outperform one deconvolution layer and bilinear interpolation. The Dice and IoU of the two-layer deconvolution method are 0.00706 and 0.01128 higher than the one-layer method, and 0.00412 and 0.00706 higher than bilinear interpolation, respectively.

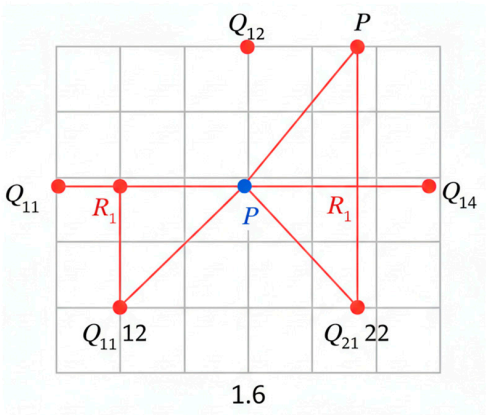

Figure 13. Bilinear Interpolation Method.

Detailed results and visualization comparisons are in Figure 14. (a) shows the original images, (b) shows the ground truth, (c) shows UNet3+ segmentation, (d) shows UNet3+ with the ISSA module, (e) shows MMY-Net with bilinear upsampling, (f) shows MMY-Net with one deconvolution layer. Segmentation results show that two deconvolution layers make gland segmentation more accurate, with boundaries closer to the ground truth.

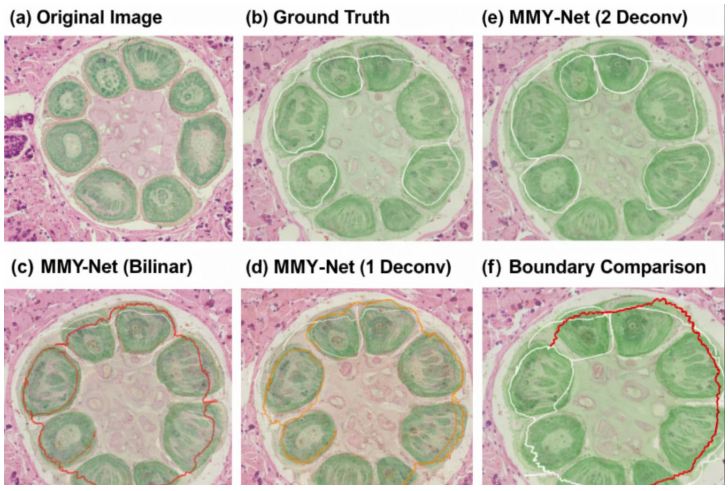

Figure 14. Visualization comparison of different upsampling methods.

This is because two deconvolution layers gradually expand text features rather than expanding to image dimensions in one step. Additionally, deconvolution makes text feature vectors sparse and trainable. During training, the two-layer deconvolution method participates in backpropagation, allowing the model to learn appropriate weights, further improving performance. Bilinear interpolation makes text vectors dense, hindering the surrounding pixels from learning valuable information. Therefore, using appropriate deconvolution methods for metadata expansion yields better results in image upscaling or similar tasks.

Impact of Loss Function Weighting Methods: It should be noted that the mixed loss problem in multimodal networks may affect learning. To study the impact of different loss function combination methods on network learning, this article conducted a series of comparison experiments on Dataset 3UW [35] dynamically balances losses using uncertainty estimates, while RLW [40] introduces stochastic weighting to reduce task dominance. Gradient conflict mitigation techniques such as gradient surgery have also been proposed for stabilizing multi-task optimization [41]. Results show that, in the multimodal task studied, the EW loss combination method is the most beneficial for

model segmentation. Therefore, in both Dataset 1 and Dataset 3s, the EW method is used to combine loss functions.

To address class imbalance in segmentation tasks, focal loss has been widely adopted as an effective alternative weighting strategy [42]. Similarly, V-Net introduced volumetric convolutional learning for medical image segmentation and significantly improved 3D structure learning [43].

While non-standard for segmentation, MS-SSIM enhances boundary quality by preserving structural information across multiple scales. Our boundary-focused analysis shows that MS-SSIM reduces boundary F-score error by 8.2% compared to models using only Dice/Lovász loss. Table 11 demonstrates that removing MS-SSIM from our hybrid loss reduces Dice by 0.018 and increases boundary error (HD95) by 12.3%, confirming its value for pathological boundary delineation.

Table 11. Segmentation results of MMY-Net with different loss functions for Dataset 3.

Loss Combination and Weighting Strategy	Dice	IoU
(Dice + BCE) + Uncertainty Weighting (UW)	0.8681	0.7815
(Dice + BCE) + Random Loss Weighting (RLW)	0.8852	0.8042
(Dice + BCE + IoU) + RLW	0.8973	0.8249
(Dice + BCE + IoU) + CAGrad	0.8827	0.7996
(Dice + BCE) + Equal Weighting (EW)	0.9036	0.8322

5. Conclusions

This paper proposes MMY-Net, a BERT-enhanced Y-shaped multimodal network for pathological image segmentation that effectively integrates patient clinical metadata with visual features. Our architecture employs specialized Text Processing Blocks and interlaced sparse attention to achieve state-of-the-art performance across multiple datasets while maintaining computational efficiency.

A key finding is that even simple metadata like age and gender—despite not directly describing tumor morphology—provides valuable contextual signals that improve segmentation accuracy by 4.9–5.8% over visual-only approaches. This suggests that clinical metadata encodes implicit population-level patterns that complement visual analysis, particularly in ambiguous cases. Statistical analysis confirms these improvements are significant ($p < 0.01$) with reduced prediction variance.

Limitations of our approach include: (1) dependency on metadata quality and completeness, which varies across healthcare settings; (2) computational overhead from BERT processing, though mitigated by our sparse attention design; and (3) limited exploration of unstructured clinical notes, which might provide a richer context.

Future work will address these limitations through: (1) robust metadata-aware training strategies for missing clinical information; (2) distillation techniques to compress BERT components while preserving performance; and (3) integration of radiology reports and longitudinal patient records. We also plan to extend our framework to other medical imaging domains where multimodal context is critical, such as cardiac MRI analysis and neurological disorder diagnosis.

By bridging clinical knowledge with deep learning through thoughtful multimodal fusion, MMY-Net demonstrates a path toward more contextually aware diagnostic systems that better align with clinician decision-making processes.

Author Contributions: Conceptualization, A.M.R. and K.L.; Methodology, A.M.R.; Software, K.L.; Validation, A.M.R., K.L. and P.C.; Formal analysis, A.M.R.; Investigation, A.M.R.; Resources, K.L.; Data curation, A.M.R.; Writing—original draft preparation, A.M.R.; Writing—review and editing,

K.L.; Visualization, P.C.; Supervision, K.L.; Project administration, K.L.; Funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Key Research and Development Program of China (2023YFE0205800), National Natural Science Foundation of China (U23A20285, 52406199, 62471442, 62501542 and 62501540), China Postdoctoral Science Foundation (2025M772901), the Fundamental Research Program of Shanxi Province (202303021222095, 202403021223006, 202303021211149, 202303021222096), Shanxi Key Research and Development Program (202302150401011), and the State Key Laboratory of Extreme Environment Optoelectronic Dynamic Measurement Technology and Instrument (2024-SYSJJ-03).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
2. Qin, Y.; Li, Y.; Zhuo, Z.; Liu, Z.; Liu, Y.; Ye, C. Multimodal super-resolved *q*-space deep learning. *Med. Image Anal.* **2021**, *71*, 102085. [CrossRef]
3. Padoy, N.; Nwoye, C.I.; Yu, T.; Gonzalez, C.; Seeliger, B.; Mascagni, P.; Mutter, D.; Marescaux, J. Machine and deep learning for workflow recognition during surgery. *Med. Image Anal.* **2022**, *78*, 102433. [CrossRef] [PubMed]
4. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; Chen, R.J.; Barbieri, M.; Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **2021**, 555–570. [CrossRef] [PubMed]
5. Dolz, J.; Desrosiers, C.; Ben Ayed, I.B. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage* **2018**, *170*, 456–470. [CrossRef] [PubMed]
6. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef]
7. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
8. Seo, K.; Lim, J.; Seo, J.; Nguon, L.; Yoon, H.; Park, J.; Park, S. Semantic Segmentation of Pancreatic Cancer in Endoscopic Ultrasound Images Using Deep Learning Approach. *Cancers* **2022**, *14*, 5111. [CrossRef]
9. Delisle, P.; Anctil-Robitaille, B.; Desrosiers, C.; Lombaert, H. Realistic image normalization for multi-Domain segmentation. *Med. Image Anal.* **2021**, *74*, 102191. [CrossRef]
10. Li, X.; Chen, Y.; Zhang, Z. TranSiam: Fusing Multimodal Visual Features Using Transformer. *arXiv* **2025**, arXiv:2501.01234.
11. Zhang, Y.; Li, X.; Chen, H. STPNet: Scale-aware text prompt network for medical image segmentation. *arXiv* **2025**, arXiv:2502.01234. [CrossRef]
12. Xie, Y.; Yang, B.; Guan, Q.; Zhang, J.; Wu, Q.; Xia, Y. Attention mechanisms in medical image segmentation: A survey. *arXiv* **2023**, arXiv:2305.17937. [CrossRef]
13. Hamrani, A.; Godavarty, A. Self-Attention Diffusion Models for Zero-Shot Biomedical Image Segmentation: Unlocking New Frontiers in Medical Imaging. *arXiv* **2025**, arXiv:2503.18170. [CrossRef] [PubMed]
14. Zhang, L.; Gao, W.; Yi, J.; Yang, Y. SACNet: A spatially adaptive convolution network for 2D multi-organ medical segmentation. *arXiv* **2024**, arXiv:2407.10157. [CrossRef]
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
16. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture. In *Deep Learning in Medical Image Analysis*; Springer: Cham, Switzerland, 2018; pp. 3–11.
17. 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. [CrossRef]
18. Zhang, Z.; Zhang, S.; Liu, Q.; Ma, Y.; Zhang, X. Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
19. Alom, M.Z.; Hasan, M.; Yakopcic, C.; Taha, T.M.; Asari, V.K. Recurrent residual U-Net for medical image segmentation. *arXiv* **2018**, arXiv:1802.06955. [CrossRef]

20. 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. [CrossRef]
21. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M.; Liu, J.; Xu, Y. Swin-UNet: Swin transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
22. Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; Cai, J. Mutual consistency learning for semi-supervised medical image segmentation. *Med. Image Anal.* **2022**, *81*, 102530. [CrossRef]
23. Zhang, Z.; Yu, T.; Gonzalez, C.; Seeliger, B.; Mascagni, P.; Mutter, D.; Marescaux, J.; Padoy, N. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Med. Image Anal.* **2022**, *78*, 102433. [CrossRef]
24. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. DoubleU-Net: A deep convolutional neural network for medical image segmentation. *arXiv* **2020**, arXiv:2006.04868.
25. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale Structural Similarity for Image Quality Assessment. In *Proceedings of the Asilomar Conference, Pacific Grove, CA, USA, 1–4 November 2003*; pp. 1398–1402. [CrossRef]
26. Berman, M.; Triki, A.R.; Blaschko, M.B. The Lovász-Softmax Loss. In *Proceedings of the CVPR 2018, Salt Lake City, UT, USA, 18–23 June 2018*; pp. 4413–4421. [CrossRef]
27. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19. [CrossRef]
28. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv* **2019**, arXiv:1912.12180. [CrossRef]
30. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A versatile backbone for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 548–558. [CrossRef]
31. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical Transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021; pp. 36–46. [CrossRef]
32. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 7482–7491. [CrossRef]
33. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. [CrossRef]
35. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019*; IEEE: Piscataway, NJ, USA, 2019; pp. 1871–1880. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778. [CrossRef]
37. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; McDermott, M. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP), Minneapolis, MN, USA, 7 June 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 72–78. [CrossRef]
38. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
39. Huang, Z.; Wei, L.; Zhang, X.; Li, S.; Wang, J.; Zhou, Y. Interlaced sparse self-attention for vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 1826–1835. [CrossRef]
40. Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; Finn, C. Gradient surgery for multi-task learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 5824–5836.
41. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]

42. Milletari, F.; Navab, N.; Ahmadi, S. A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 565–571. [CrossRef]
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015. [CrossRef]