Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid U-Net Model with Transformer-Enhanced Skip Connections for Abdominal Organ Segmentation in Laparoscopic Images

Ashok Kumar

ashokshravasti@gmail.com

Jawaharlal Nehru University

Satish Chand

Jawaharlal Nehru University

Research Article

Keywords: Organ Segmentation, Surgical Images, Laparoscopic-cholecystectomy, Vision Transformer, U-Net

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10863919/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid U-Net Model with Transformer-Enhanced Skip Connections for Abdominal Organ Segmentation in Laparoscopic Images

Ashok Kumar[1*] and Satish Chand[1††]

[1*]School of Computer and Systems Sciences, Jawaharlal Nehru University, 110067, New Delhi, India.

*Corresponding author(s). E-mail(s): ashokshravasti@gmail.com;
Contributing authors: schand20@gmail.com;
†These authors contributed equally to this work.

Abstract

The organ Segmentation in surgical images plays a crucial role in computer-assisted diagnosis and robotic surgeries, yet it faces challenges like class imbalance, occlusions, and computational constraints. In order to overcome these challenges, we propose an architecture which uses asymmetric multi-scale transformer along with the attention-gated skip connections in U-net Model to efficiently segment abdominal organs. Our proposed architecture allows efficient extraction of local features and modelling of global contexts. We also propose a hybrid loss function by combining Lovasz-Softmax loss and sparse categorical cross-entropy loss, which helps in overcoming class imbalance in the training process and improving performance on the minor classes. Our proposed method attains a mean precision of 0.9667, mean recall of 0.9432, mean dice of 0.9522, mean IoU of 0.9150, and mean HD95 of 0.4704 on all 13 classes of the CholecSeg8k dataset, which makes it better than the state-of-the-art techniques with low computational cost.

Keywords: Organ Segmentation, Surgical Images, Laparoscopic-cholecystectomy, Vision Transformer, U-Net

1 Introduction

Image segmentation is a crucial job in the field of computer vision that partitions an image into useful segments that are useful for subsequent processing and understanding. Image segmentation plays a very important role in medical applications by delineating various anatomical or pathological structures within images acquired from techniques such as MRI, CT, ultrasound, and surgical imaging. Among these tasks, organ segmentation is particularly important for understanding structural boundaries and spatial relationships within the human body. While organ segmentation has been extensively studied in structured imaging modalities such as CT and MRI, increasing attention has been directed toward surgical imaging, where visual data is acquired through endoscopic, laparoscopic and other systems. In these settings, segmentation is required to identify organs and anatomical landmarks. A prominent example is laparoscopic cholecystectomy, which remains one of very popular minimally invasive procedures, where accurate segmentation of organs such as the liver, gallbladder, and surrounding structures is essential for safe and effective surgical intervention. Surgical image segmentation presents a particularly challenging problem, as surgical scenes are characterised by dynamic environments, occlusions, illumination variations, and the presence of deformable tissues and instruments[1]. These factors, combined with the presence of small, low-contrast anatomical structures, make reliable segmentation significantly more difficult[2]. Accurate segmentation in such settings is crucial for computer-assisted and robotic-assisted surgery, where reliable identification of anatomical structures and surgical tools directly supports intraoperative guidance, workflow analysis, and context-aware decision-making [3, 4]. Furthermore, precise segmentation of surgical images assists clinicians in preoperative planning and enhances the safety and efficiency of surgical procedures. Medical professionals use Classical approaches, such as thresholding[5], region growing[6], edge detection[7], and deformable models[8], that rely heavily on handcrafted features and predefined assumptions about image intensity, texture, or shape. These methods are time-consuming, highly sensitive to noise, illumination variations, and structural complexity, making them unsuitable for the highly dynamic and unstructured nature of surgical scenes. To overcome these limitations, machine learning-based techniques were introduced, where features are learned from data rather than manually designed. Support Vector Machines and random forests are ML-based classifiers applied to medical image segmentation tasks by leveraging hand-crafted features [9]. These approaches improve generalization compared to traditional methods; however, their performance remains constrained due to the quality and expressiveness of the handcrafted features, which often fail to capture complex spatial and contextual relationships present in surgical images.

The emergence of deep learning has significantly changed medical image segmentation by enabling end-to-end feature learning directly from raw data. Convolutional neural networks (CNNs) have exhibited a resilient capacity to learn hierarchical representations that include both low-level and high-level information [10]. Building upon this, a specialised encoder–decoder architecture called U-Net [11] is introduced for biomedical image segmentation. Its design combines a contracting path for context extraction with an expanding path for precise localization, and skip connections facilitate the transfer of high-resolution spatial information from the encoder to the decoder. Through the

skip connections, the U-Net model recovers spatial features lost during pooling operations in the encoder phase. Despite their effectiveness, skip connections in conventional U-Net are implemented as direct feature transfer mechanisms, where encoder features are concatenated with decoder features without considering their contextual relevance. In complex surgical scenes, such as those in the CholecSeg8k dataset, this design becomes problematic due to the presence of small, low-contrast, and highly deformable structures, leading to severe class imbalance and boundary ambiguity. Under these conditions, skip connections often propagate ambiguous or noisy representations, limiting the effectiveness of feature fusion during decoding. The immediate transfer of attributes from the encoder to the decoder without acknowledging these complexities can widen the semantic gap, thereby restricting the model's effectiveness. Existing works have attempted to address this limitation by modifying skip connections through dense pathways or additional convolutional transformations to minimize the semantic gap between the features of the encoder and decoder [12–15]. However, these methods mostly concentrate on structural connectivity and local transformations, and remain limited in their ability to selectively refine and filter the information being transferred. These limitations indicate the need for a mechanism that can refine the information propagated through skip connections in a context-aware manner, without increasing the complexity of the encoder–decoder backbone.

In this work, we revisit the role of skip connections and propose a two-stage skip connection refinement strategy for organ segmentation in the CholecSeg8k dataset. Specifically, encoder features at two deeper levels, which contain more fine-grained details, are first enhanced using transformer-based modules to incorporate global contextual relationships and reduce feature ambiguity. These refined features are then passed through attention-gated mechanisms, enabling selective suppression of irrelevant activations and emphasizing task-relevant information before fusion. Subsequently, the decoder receives the refined features for segmentation and resolution reconstruction. This sequential refinement context enhancement using multiscale asymmetric transformers, followed by attention-based filtering ensures that only informative and context-aware features are propagated through skip connections. This leads to more effective feature integration and improved segmentation performance in challenging surgical scenarios. The proposed architecture is further optimised using a hybrid loss formulation that combines Lovász-Softmax loss [16] with Sparse Categorical Cross-Entropy loss. In multi-class surgical scene segmentation, classical pixel-wise loss functions such as Cross-Entropy primarily optimise local classification accuracy and pixel-level discrimination, but often fail to directly optimise region-based evaluation metrics such as IoU and HD95, especially when there is a significant class imbalance, as in the CholecSeg8k dataset. This limitation becomes especially critical for small, thin, and low-contrast anatomical features like the liver, cystic ducts, and blood vessels, which occupy only a sparse set of pixels and are highly susceptible to boundary ambiguity and misclassification. Proposed hybrid loss addresses these challenges by integrating pixel-level supervision with overlap-based structural optimisation. It improves boundary preservation, enhances sensitivity toward minor classes,

mitigates bias toward dominant anatomical regions, and promotes more balanced feature learning across both large and underrepresented structures.
In summary, our contributions are as follows:

- We employ a **selective transformer placement strategy** to apply refinement of skip connection features only at deeper levels while preserving high-resolution features.
- We have proposed a hybrid loss function that integrates the Lovász-Softmax loss with sparse categorical cross-entropy to address class imbalance and improve the segmentation performance of small and ambiguous structures.
- The performance of the proposed model is compared with six state-of-the-art transformer and convolution-based models.

Overall, this work revisits the role of skip connections as an information-flow mechanism rather than a simple feature-transfer pathway. By refining skip features through context enhancement and attention-based filtering, while preserving a purely convolutional encoder–bottleneck backbone, the model is the right choice for real-time surgical image analysis since it achieves better segmentation performance with less processing overhead.

2 Related Work

In this section, we review the literature related to organ segmentation. Organ segmentation has attracted researchers due to its various applications, especially in medical science. Early traditional segmentation methods relied on basic color histogram thresholding and region growing to map structures like the colonic lumen [17], but these struggled with intraoperative lighting artefacts. To address this, Arnold et al.[18] introduced color-space thresholding to specifically segment and inpaint specular highlights, though basic thresholding still mischaracterized complex tissue reflections. Seeking more robust structural mapping, researchers shifted to active contours, where Jimenez-Carretero et al. [19] overcame the classical level-set limitation of "leaking" into adjacent tissues by applying local curvature constraints. Still, capturing glare on moist tissues remained difficult, prompting Marcinczak and Grigat[20] to improve upon earlier basic thresholding methods by utilising closed-contour morphological operations to better isolate reflections. Ultimately, acknowledging that all purely intraoperative image-based methods fail under severe blood occlusion and poor lighting, Nosrati et al.[21] addressed these collective visibility limitations by anchoring their multi-organ level-set contours with pre-operative 3D anatomical models. These methods lack robustness to surgical variations such as motion, lighting, deformation, high manual intervention, and poor generalisation across patients or procedures. Classical machine learning methods, such as Support Vector Machines[22], Random Forests[23], enabled learning from data without deep hierarchical representations. They offered better generalisation than traditional methods by learning discriminative features and handling multi-class problems, yet still relied heavily on manual feature engineering and struggled with complex contextual relationships and high intraoperative image variability [24].

2.1 Medical Image Segmentation with U-Net and its Variants

Introduction of CNN-based methods, such as Full Convolutional Networks for Automatic hierarchical feature learning, end-to-end training, mitigates the problem in the ML-based methods, which are excellent on local details and accurate boundary segmentation via skip encoder. The U-Net architecture, introduced by [25], has become the de facto standard for biomedical image segmentation due to its symmetric encoder-decoder structure with skip connections. The paper [26] proposed a modified U-Net model for the segmentation of surgical instruments and operated anatomical structures, employing two decoder branches to enhance feature learning and precision in complex environments. The study[27] is one of the earliest comprehensive comparisons of U-Net, TernausNet (U-Net with pretrained VGG encoders), LinkNet, FCN, and SegNet on laparoscopic cholecystectomy images for multi-class segmentation (liver, gallbladder, instruments, fat, etc.). Among these models TernausNet with soft-Jaccard loss achieved strong results up to 0.78 IoU. In paper [28], researchers have employed a basic U-Net model with the SE-RESNET50 Model as a backbone for feature extraction to segment the liver. Previous work on this dataset has established strong baselines using models like U-Net and its variants. Various modifications have been proposed to improve the effectiveness of the conventional U-Net. UNet++ [29], which incorporates nested and dense skip connections to minimize the semantic gap between the feature maps of the encoder and decoder. Similarly, [30] integrated attention gates into the U-Net design, allowing the model to suppress unimportant areas and emphasise key structures, which is particularly beneficial in noisy surgical environments. Architectures like MobileNet [31] and ShuffleNet [32] were designed specifically for efficiency. They significantly reduce the number of parameters and floating-point operations (FLOPs) while retaining competitive performance by using depthwise separable convolutions and channel shuffling. These networks have been successfully adopted as efficient encoders (backbones) in segmentation models like Fast-SCNN [33] and Lite-RASSP [34]. CCNN-based techniques have small receptive fields and have issues in extracting global context and handling variable organ shapes/deformations, leading to overfitting on small datasets. CNN locality bias affects performance on highly variable surgical scenes[35].

2.2 Transformers and Hybrid Architectures in Medical Imaging

The Transformer architecture [36], renowned for its success in natural language processing, has recently been adapted for computer vision tasks as Vision Transformers (ViTs) [37]. Subsequently, researchers have explored integrating Transformers into medical image segmentation. In the study [38], researchers proposed TransUNet, which demonstrates outstanding results by combining a CNN for extracting features with a Transformer for global context retrieval. Similarly, [39] introduced MedT, which uses a gated axial-attention mechanism for medical image segmentation. However, these pioneering models are often computationally intensive as self-attention is quadratic in complexity, resulting in them being less suitable for settings with limited resources.

The inability of the U-Net to capture long-range dependencies is addressed by the U-Transformer by incorporating attention at multiple levels. It employs self-attention to model global context in encoder features and cross-attention within skip connections to refine spatial details by suppressing irrelevant features on abdominal CT datasets [40].In paper[41], researchers combined Swin Transformer's window-based attention with Deformable Transformers for adaptive sampling, improving convergence and alignment with object shapes. UCTransNet[12] presents the CTrans module as an alternative to conventional skip connections in order to reduce the discrepancy between encoder and decoder features. This approach utilizes Channel-wise Cross-fusion with Transformer (CCT) and Channel-wise Cross-Attention (CCA) to adaptively compensate for semantic gaps across different scales[12]. U-Net v2 improves medical image segmentation by employing a novel skip connection mechanism that infuses semantic information into low-level features via the Hadamard product[42]. By reviewing this literature, we found that employing a transformer to enhance the skip connection in the U-Net model will improve model performance on minor classes, strengthen global relationships, and mitigate semantic differences in the encoder and decoder.

3 Proposed Model

In this section, we introduce our proposed hybrid model for organ segmentation in laparoscopic images, which consists of a U-Net model, comprising an encoder, a bottleneck, and decoder blocks. The encoder block extracts the features from the input, the bottleneck reduces the feature map obtained from the encoder and the decoder generates the segmentation map. Further, we propose a hybrid loss function that helps in identifying the small anatomical structure in the dataset.

The encoder has 4 stages, each comprising 2 convolution operations, batch normalisation, and a ReLU Activation function, and produces skip features denoted as S_l. The skip feature undergoes the max-pooling operation, producing the pooled features (feature map), denoted as P_l. Consider an input image P_0 with spatial dimension $H \times W \times C$, the S_l and P_l are computed as follows:
$S_1 = \mathcal{F}(P_0) \qquad P_1 = \text{MaxPool}(S_1)$
$S_2 = \mathcal{F}(P_1) \qquad P_2 = \text{MaxPool}(S_2)$
$S_3 = \mathcal{F}(P_2) \qquad P_3 = \text{MaxPool}(S_3)$
$S_4 = \mathcal{F}(P_3) \qquad P_4 = \text{MaxPool}(S_4)$
where $\mathcal{F}$ denotes two consecutive Conv–BN–ReLU units, with the number of filters specified per stage.The above operations can be written in compact form as:

$$S_l = \mathcal{F}(P_{l-1}), \qquad P_l = \text{MaxPool}(S_l), \qquad l = 1, 2, 3, 4, \tag{1}$$

$P4$ is passed into the bottleneck block, where the depth of the representation is maximized while its spatial size remains unchanged. For this, the same operations used in the encoder block except the maxpooling are applied on the feature map P_4. Feature depth maximisation increases the number of channels, which in turn maintains the features specially related to small organs like, liver etc. As a result, it identifies abstract concepts like edges in the organ. The resultant feature map, also called global

features, denoted by B, from the bottleneck is obtained as follows:

$$\mathcal{B} = \mathcal{F}(P_4), \tag{2}$$

Now we refine the skip connection features: S_3 and S_4, using the transformer(introduced by [36]) to model global context. A 1×1 convolutional operation $\phi_l(\cdot)$ projects each selected feature map, which is then flattened into a token sequence as follows:

$$Z_l = \text{Flatten}\big(\phi_l(S_l)\big) \in \mathbb{R}^{N_l \times d_l}, \qquad l \in \{3, 4\}, \tag{3}$$

where $N_l = H_l W_l$ is the number of tokens and d_l the embedding dimension. The spatial arrangement of the tokens is retained by incorporating trainable position embeddings $E_l \in \mathbb{R}^{N_l \times d_l}$ into the flattened features Z_l, denoting as $\hat{Z}_l$:

$$\hat{Z}_l = Z_l + E_l. \tag{4}$$

The tokens $\hat{Z}_l$ are passed through layer normalisation (LN) to obtain the normalised tokens, which are passed through the MHSA, to captures long-range dependencies. The MHSA takes the normalised tokens ($\text{LN}(\hat{Z}_l)$), splits them into h heads, and computes attention in each head independently as follows:

$$\text{MHSA}((LN(\hat{Z}_l)) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O, \qquad \text{head}_i = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \tag{5}$$

where $Q_i = LN(\hat{Z}_l)W_i^Q$ is query vector, $K_i = LN(\hat{Z}_l)W_i^K$ is key vector, $V_i = LN(\hat{Z}_l)W_i^V$ is the value vector, and $d_k = d_l/h$ is the per-head dimension. Adding the input of a layer to it's output is termed as residual connection and it's process is shown below:

$$Z_l^{(1)} = \hat{Z}_l + \text{MHSA}\big(\text{LN}(\hat{Z}_l)\big), \tag{6}$$

The residual connection preserves the original token features, and provides a direct gradient path that stabilises training. Further, the resultant tokens $Z_l^{(1)}$ are normalized and then passes through the feed-forward MLP (using GELU activation function), to refine the token features as follows:

$$Z_l^{(2)} = \text{GELU}((\text{LN}(Z_l^{(1)})\, W_1 + b_1)\, W_2 + b_2, \tag{7}$$

where $W_1 \in \mathbb{R}^{d_l \times 2d_l}$ and $W_2 \in \mathbb{R}^{2d_l \times d_l}$ are the weight matrices optimized during training for the two feed-forward layers. We now apply the residual connection on the MLP processed tokens as follows:

$$Z_l^{(out)} = Z_l^{(1)} + Z_l^{(2)}, \tag{8}$$

The complete sequence of operations- Layer Normalization, MHSA, and a residual addition, followed by Layer Normalization, the MLP, and a second residual addition constitutes a single transformer block, denoted by $\mathcal{T}(\cdot)$. In order to obtain global

context from skip connection features S_3 and S_4; S_3 is passed through a single transformer block while S_4 is passed through the transformer block successively two times as follows:

$$Z_3^{(\text{out})} = \mathcal{T}(\hat{Z}_3), \qquad Z_4^{(\text{out})} = \mathcal{T}\big(\mathcal{T}(\hat{Z}_4)\big). \tag{9}$$

Finally, the token sequences $Z_3^{(out)}$ and $Z_4^{(out)}$ are reshaped back into two-dimensional spatial feature maps so that they can be processed by the convolutional layers of the decoder. A 1×1 convolution $\phi_l'(\cdot)$ is then applied on this to align it with the original channel dimension of the corresponding skip feature. The refined skip feature is computed as:

$$\tilde{S}_l = \phi_l'\big(\text{Reshape}(Z_l^{(out)})\big), \qquad l \in \{3, 4\}. \tag{10}$$

Thus, we have obtained refined skip connection features $\tilde{S}_3$ and $\tilde{S}_4$. The arhchitecture diagram of the transformer block used to refine these two skip connection is presented in the Fig. 1. We also have original skip features S_1 and S_2. To make the notation

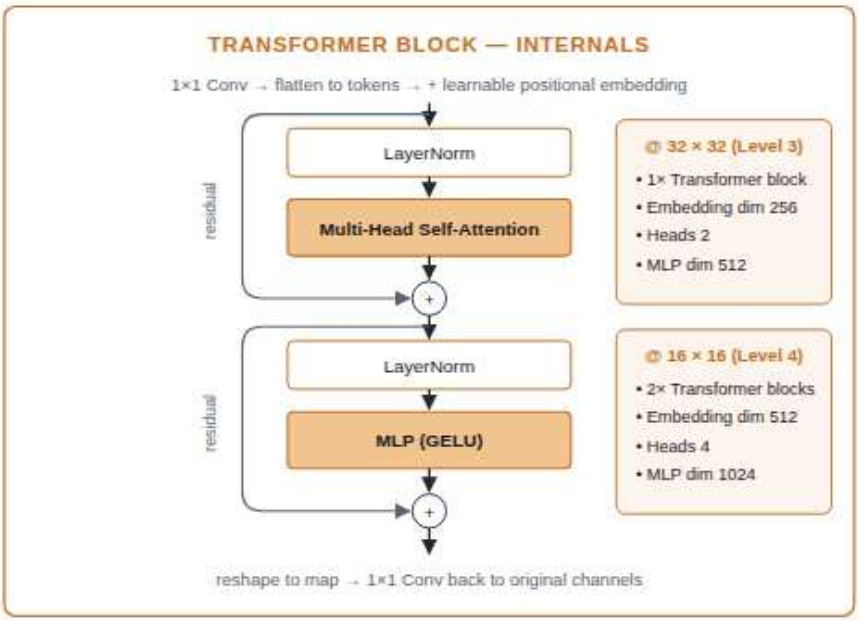


Fig. 1: Transformer Block.

uniform , we rename S_1 and S_2 as $\tilde{S}_1$, $\tilde{S}_2$ respectively. These features $\{\tilde{S}_l\}_{l=1}^4$ along with bottleneck features $\mathcal{B}$ (from equation (2)) are passed to the decoder in reverse order, ($\tilde{S}_4$,..., $\tilde{S}_1$). The decoder is composed of four phases: during the initial stage, the bottleneck representation $\mathcal{B}$ is upsampled and then an attention map is computed using upsampled feature and corresponding skip feature $\tilde{S}_4$, denoting the output

(attention map) α_4, as follows:

$$\alpha_4 = \sigma\Big(W_\psi\big(\delta(W_g\,\mathrm{Up}(\mathcal{B}) + W_x\,\tilde{S}_4)\big)\Big) \tag{11}$$

Where W_g, W_x, and W_ψ are the learnable 1×1 convolutional projections, $\delta(\cdot)$ and $\sigma(\cdot)$ denote the ReLU and sigmoid activation functions, respectively. To improve regularization, a dropout layer is first applied to the attention map α_4, followed by an element-wise multiplication($\odot$) with the skip features $\tilde{S}_4$ is performed as follows:

$$\hat{S}_4 = \mathrm{Dropout}(\alpha) \odot \tilde{S}_4, \tag{12}$$

The resultant output $\hat{S}_4$, also called as attention-refined skip feature, is concatenated with the up-sampled bottleneck feature $\mathcal{B}$. This concatenated feature is further refined by two consecutive convolution operations, each followed by batch normalization and the ReLU activation function denoted as $\mathcal{R}(\cdot)$ to produce the decoder feature D_4 as follows:

$$D_4 = \mathcal{R}\Big(\mathrm{Concat}(\mathrm{Up}(\mathcal{B}), \hat{S}_4)\Big) \tag{13}$$

The D_4 is passed to the second stage that performs same steps as the first phase, in which the bottleneck feature $\mathcal{B}$ is replaced by D_4 and skip connection feature $\tilde{S}_4$ is replaced by $\tilde{S}_3$. The resultant output, denoted as D_3, is obtained as follows:

$$D_3 = \mathcal{R}\Big(\mathrm{Concat}(\mathrm{Up}(D_4), \mathrm{Dropout}(\sigma\Big(W_\psi\big(\delta(W_g\,\mathrm{Up}(D_4) + W_x\,\tilde{S}_3)\big)\Big)) \odot \tilde{S}_3)\Big) \tag{14}$$

For the third stage, the D_4 is replaced by D_3 and $\tilde{S}_3$ is replaced by $\tilde{S}_2$ to obtain the resultant output as D_2; and for last stage, D_3 is replaced by D_2 and $\tilde{S}_2$ by $\tilde{S}_1$ to obtain final segmentation map D_1, as follows:

$$D_2 = \mathcal{R}\Big(\mathrm{Concat}(\mathrm{Up}(D_3), \mathrm{Dropout}(\sigma\Big(W_\psi\big(\delta(W_g\,\mathrm{Up}(D_3) + W_x\,\tilde{S}_2)\big)\Big)) \odot \tilde{S}_2)\Big) \tag{15}$$

$$D_1 = \mathcal{R}\Big(\mathrm{Concat}(\mathrm{Up}(D_2), \mathrm{Dropout}(\sigma\Big(W_\psi\big(\delta(W_g\,\mathrm{Up}(D_2) + W_x\,\tilde{S}_1)\big)\Big)) \odot \tilde{S}_1)\Big) \tag{16}$$

The detailed architecture diagram of the attention gate is presented in Fig. 2. The final segmentation map $D_1$is further passes through the 1×1 convolution followed by softmax activation function to get the original segmentation image as follows:

$$Y = \mathrm{Softmax}(W_{\mathrm{out}} * D_1)\,, \tag{17}$$

where W_{out} is the learnable 1×1 convolution kernel, and $Y \in \mathbb{R}^{H_0 \times W_0 \times C}$ denotes the predicted segmentation image.

In this section, we have discussed the proposed hybrid model that combines U-Net and transformer, and its architecture is shown in Fig.3. The convolutional encoder captures fine local details and spatial structure, while the transformer captures the long-range dependencies from the input image. The decoder uses the attention gates to focus on the most relevant regions while suppressing irrelevant background, and

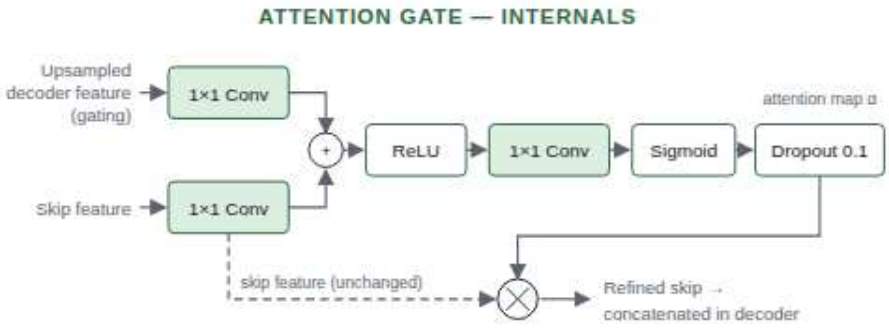


Fig. 2: Attention Gate.

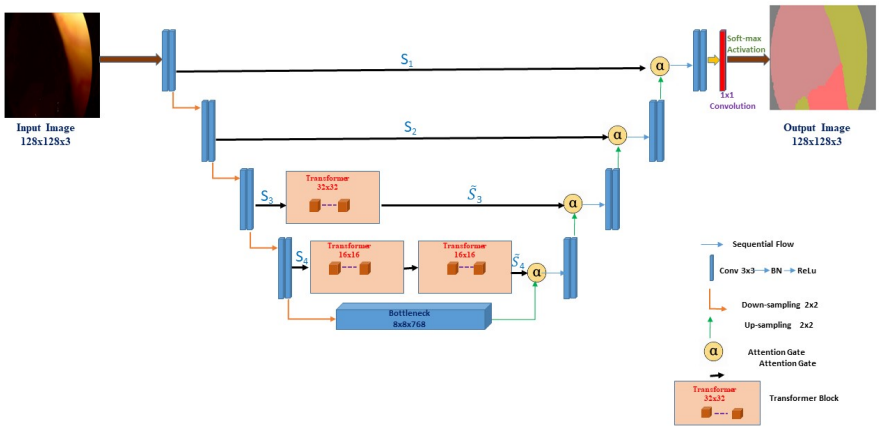


Fig. 3: Architecture of the proposed hybrid-unet model with encoder, decoder and transformer-enhanced attention-gate-based skip connections.

restores the full spatial resolution to produce the final segmentation map. The proposed model enables accurate multi-class organ segmentation in laparoscopic images. We now discuss our proposed hybrid loss function.

3.1 Hybrid Loss Function

In multi-class medical image segmentation, there exists a discrepancy between the optimization objective and the evaluation metrics because loss functions such as Cross-Entropy (CE) are used for training most segmentation networks, which optimize per-pixel classification accuracy. However, segmentation performance is typically evaluated using region-based overlap metrics such as Intersection over Union (IoU),

Dice score and Hausdorff Distance (HD95). This mismatch often leads to suboptimal region-level performance, particularly for small anatomical structures and highly imbalanced classes. To address this limitation, we introduce a hybrid loss formulation that integrates the Lovász-Softmax loss, enabling direct optimisation of a differentiable surrogate of IoU, with Sparse Categorical Cross-Entropy, which provides stable pixel-wise supervision. For a multi-class segmentation problem which has C classes, let $p_{i,c}$ represent the predicted probability that pixel i is assigned to class c, while $y_i \in \{0, \ldots, C-1\}$ denote the corresponding ground-truth class. The Sparse Categorical Cross-Entropy loss is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{i,y_i} \tag{18}$$

Where N corresponds to the number of pixels under consideration. The Lovász-Softmax loss [16] is a convex, piecewise-linear surrogate of the Jaccard (IoU) loss that can be optimised directly by gradient descent. For each class c, a per-pixel error

$$e_i^{(c)} = \left| f_i^{(c)} - p_{i,c} \right| \tag{19}$$

is computed between the foreground indicator $f_i^{(c)} = \mathbb{1}(y_i = c)$ and the predicted probability $p_{i,c}$. These errors are sorted in decreasing order and combined with the gradient of the Lovász extension of the Jaccard loss, $g^{(c)}$, to give the class-wise loss

$$\mathcal{L}_{Lovasz}^{(c)} = \sum_{k=1}^{N} e_{(k)}^{(c)} \, g_k^{(c)}, \tag{20}$$

where $e_{(k)}^{(c)}$ denotes the $k-th$ largest error and $g_k^{(c)}$ denotes the corresponding component of the Lovász gradient. The final loss is averaged over the set of classes present in the ground truth, $\mathcal{C}'$:

$$\mathcal{L}_{Lovasz} = \frac{1}{|\mathcal{C}'|} \sum_{c \in \mathcal{C}'} \mathcal{L}_{Lovasz}^{(c)}. \tag{21}$$

To capitalise the advantages of both losses, the final hybrid loss is defined as:

$$\mathcal{L}_{Hybrid} = (1-\lambda)\mathcal{L}_{Lovasz} + \lambda \mathcal{L}_{CE}, \qquad 0 \leq \lambda \leq 1 \tag{22}$$

For $\lambda = 0$, it behaves like Lovász-Softmax loss and for $\lambda = 1$ it behaves like sparse categorical cross entropy. For $0 < \lambda < 1$, it considers both the loss functions that simultaneously enhance pixel-level classification accuracy and region-level overlap optimisation, making it particularly suitable for highly imbalanced, multi-class segmentation datasets like CholecSeg8k.

4 Results & Discussion

This section presents a detailed evaluation of the performance achieved by our proposed model. We first discuss the dataset that will be used for experimental purposes, followed by performance metrics and the results.

4.1 Dataset

The dataset that will be used is CholecSeg8k dataset, is a laparoscopic semantic segmentation dataset publicly available on Kaggle, primarily for carrying out the research in computer-assisted surgery and medical image analysis. It has been acquired from the Cholec80 dataset, which includes 80 full-length cholecystectomy videos that were captured at 25 frames per second and labelled with surgical instruments and stages of the procedure. 8,080 frames were taken from 17 typical video clips from this dataset, focusing on surgical phases relevant to organ and instrument identification. Each frame

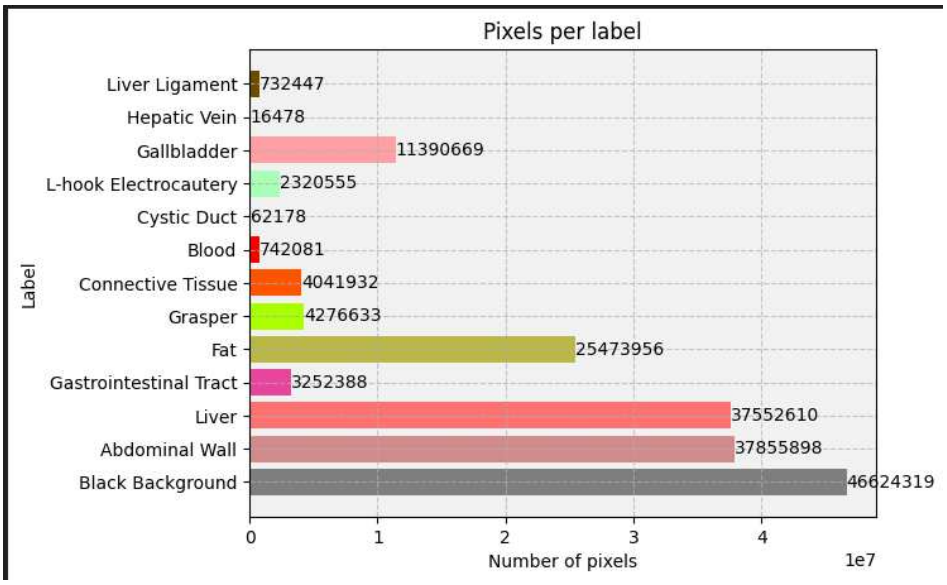


Fig. 4: Pixel count of all the 13 classes

comprises images of 854 × 480 pixels resolution along with annotations at the pixel level into 13 classes, comprising anatomical structures (abdominal wall, liver, gastrointestinal tract, fat, connective tissue, cystic duct, gallbladder, hepatic vein, and liver ligament), two surgical instruments (grasper and L-hook electrocautery), background, and blood. Table 1 summarises the pixel distribution in the dataset. As evident from the table, there is a strong imbalance across classes: large structures such as the abdominal wall and liver together occupy more than 40% of the annotated pixels, whereas thin or sparse structures such as the cystic duct (0.04%) and hepatic vein

Table 1: Distribution of pixels across the anatomical structures in the dataset, percentage-wise

Class Name	Pixel Count	Percentage (%)
Black Background	46,624,319	26.74
Abdominal Wall	37,855,898	21.71
Liver	37,552,610	21.54
Fat	25,473,956	14.62
Gallbladder	11,390,669	6.53
Grasper	4,276,633	2.45
Connective Tissue	4,041,932	2.32
Gastrointestinal Tract	3,252,388	1.87
L-hook Electrocautery	2,320,555	1.33
Blood	742,081	0.43
Liver Ligament	732,447	0.42
Cystic Duct	62,178	0.04
Hepatic Vein	16,478	0.01

Table 2: Model and training hyperparameters.

Hyperparameter	Value
Architecture	
Input size	$128 \times 128 \times 3$
Base filters	64
Transformer embedding dim (d_3, d_4)	256, 512
Attention heads (S_3, S_4)	2, 4
Attention-gate dropout	0.1
Training	
Optimiser	AdamW (weight decay 1×10^{-4})
Learning-rate schedule	Cosine decay, $1 \times 10^{-3} \rightarrow 1 \times 10^{-5}$
Batch size	16
Epochs	200
Gradient clipping	clipnorm = 1.0
λ	0.3

(0.01%) each account for well under 1%. Model and training hyperparameter used for the experiment are mentioned in Table 2:

4.2 Performance Metrics

The performance of the proposed model is quantified through a set of overlap- and boundary-based metrics, all of which are derived from the pixel-level confusion matrix accumulated over the entire test set. For a multi-class segmentation problem with C classes, the confusion matrix M is a C × C matrix in which each entry M(i, j) records the number of pixels whose ground-truth label is class i and whose predicted label is class j. The diagonal entries therefore represent correctly classified pixels, while the off-diagonal entries capture the confusions between classes. For any class c, the true

positives, false positives and false negatives are read directly from the matrix as:

$$\mathrm{TP}_c = M(c,c), \tag{23}$$

$$\mathrm{FP}_c = \sum_i M(i,c) - M(c,c), \tag{24}$$

$$\mathrm{FN}_c = \sum_j M(c,j) - M(c,c), \tag{25}$$

$$\mathrm{TN}_c = N - \mathrm{TP}_c - \mathrm{FP}_c - \mathrm{FN}_c, \tag{26}$$

that is, FP_c counts the pixels of other classes predicted as c, and FN_c counts the pixels of class c predicted as some other class. Five evaluation metrics Precision, Recall, Dice Score, Intersection over Union (IoU) based on the confusion matrix values and Hausdorff Distance (HD95) are employed to assesses the performance of proposed model and compare it with other models. These metrics provide a detailed understanding of the model's ability to correctly identify and localize various abdominal organs.

4.2.1 Precision.

Precision quantifies the model's reliability in identifying organ regions without including background or neighbouring tissues by measuring the ratio of correctly predicted positive pixels to all pixels predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{27}$$

4.2.2 Recall.

Recall reflects the model's completeness in capturing organ regions by calculating the proportion of actual organ pixels correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{28}$$

4.2.3 Dice Score.

The harmonic mean of Precision and Recall is called the Dice Score, which provides a balanced assessment that considers both false positives and false negatives. It is very useful for addressing class imbalance in organ segmentation tasks.

$$\text{Dice Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{29}$$

4.2.4 Intersection over Union (IoU).

IoU evaluates the degree of agreement between the predicted and ground-truth organ regions by calculating the overlap area divided by the combined area of both regions.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{30}$$

4.2.5 Hausdorff Distance (HD95).

HD95 assess the boundary correctness between the ground truth mask and the anticipated segmentation. It assesses spatial disparity between contour points, contrary to overlap-based metrics like IoU and Dice. The 95th percentile of the Hausdorff distance is considered to reduce sensitivity to outliers. A lower HD95 value denotes symmetric boundary alignment and more accurate object contour segmentation. Let G and P represent the ground truth and predicted segmentation's sets of boundary points, respectively. The definition of the HD95 is:

$$HD_{95}(G, P) = \max\left(\text{percentile}_{95}\{d(g, P)\},\ \text{percentile}_{95}\{d(p, G)\}\right) \tag{31}$$

where $d(g, P) = \min_{p \in P} \|g - p\|$ and $d(p, G) = \min_{g \in G} \|p - g\|$ are the smallest Euclidean distance between a point and the boundary on the other side. Precision emphasizes the model's specificity, Recall reflects its sensitivity, Dice Score balances both, IoU provides a spatial overlap measure, and HD95 checks the boundary accuracy — together offering a holistic assessment of the model's efficiency in delineating multiple organ boundaries within the CholecSeg8k dataset.

4.3 Experimental Results

The proposed model has been trained using the CholecSeg8k dataset by considering the hyperparameters, as given in Table 2 . We have used TensorFlow/Keras for model implementation and the Kaggle platform for training on a GPU P100. We have divided the dataset into 80% for training, which roughly consists of 6,464 images, 10% for validation, which roughly has 808 images, and 10% for testing the model. The images and the corresponding masks have been resized to 128 × 128 pixels, and the pixel intensities have been normalised to the range [0, 1]. To reduce the impact of class imbalance across the dataset and to enhance model generalisation, we have used seven augmentation techniques: random rotation, random horizontal flip, random vertical flip, random crop, random brightness adjustment, random contrast adjustment, and random saturation adjustment to increase the examples in the training dataset, which increases the number of images from 6,464 to 9,081. We have used the proposed hybrid loss function (already discussed in Section 3.1) as a monitor for training. To compute the training and validation losses, we have trained the model for 200 epochs, as shown in Fig. 5. The plots of training and validation loss should both decrease steadily to indicate effective learning. As evident from the Fig. 5, both losses continue to decrease up to epoch 134 and beyond that, training loss still continues to decline while the validation loss shows an upward trend. Model weights are saved at the min-

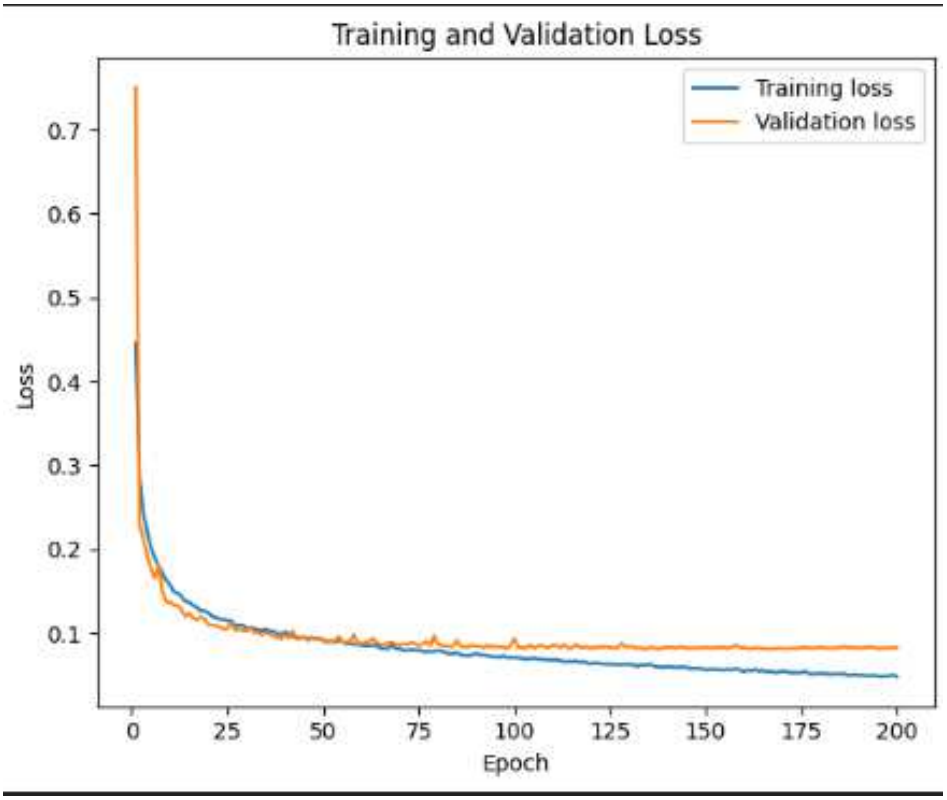


Fig. 5: Training curve of the model with training, validation loss and accuracy

imum validation loss, and to examine the model's discriminative behaviour between anatomical structures, a row-normalised confusion matrix has been computed on the test dataset, in which each diagonal entry corresponds to the per-class recall and the off-diagonal entries expose the systematic misclassifications made by the network as shown in Fig. 6.

4.3.1 Model Performance

We have also presented the result of the proposed model for all the performance parameters for individual classes, as shown in Table 3. The model obtains high performance on the large and well-defined structures such as Black Background, Abdominal Wall, Liver, Fat, and Gallbladder, with a mean Precision of 0.987, Recall of 0.989, Dice of 0.987, IoU of 0.976 and HD95 of 0.192 over these six classes. These structures have sufficient training pixels, and the convolutional encoder can learn their texture and boundaries directly. Their boundaries are also long and smooth, and therefore they are preserved through the pooling and upsampling operations, which is the reason for the low HD95 values obtained for them. For the moderately complex regions such as Gastrointestinal Tract, Grasper, Connective Tissue and L-hook Electrocautery, the

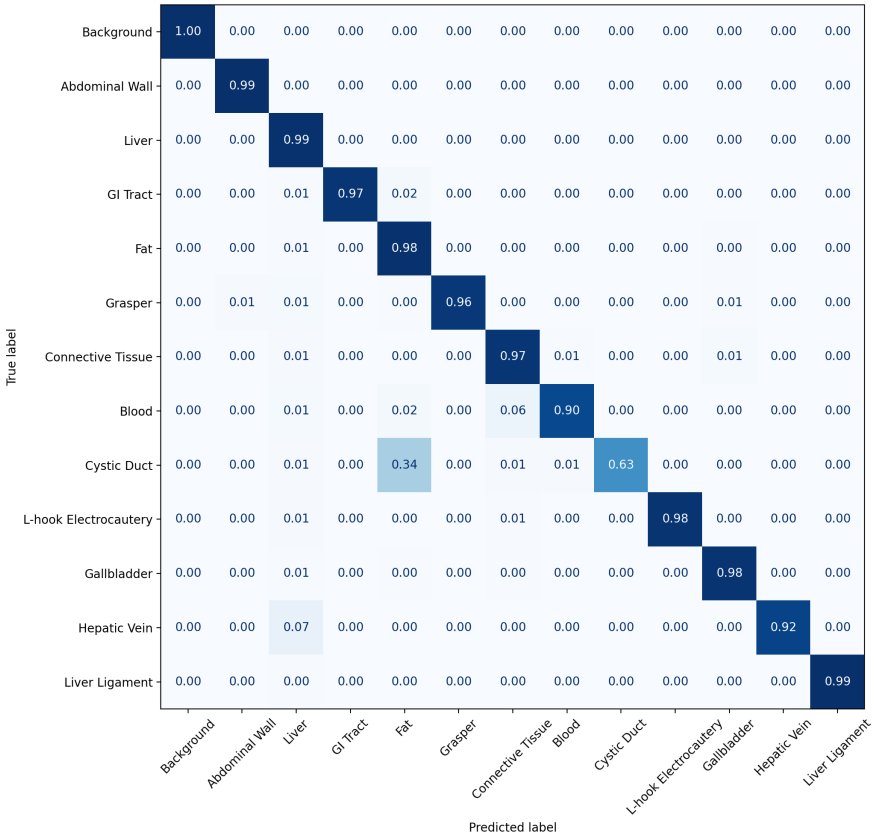


Fig. 6: Confusion Matrix over the test dataset

Table 3: Class-wise quantitative performance of the proposed Hybrid U-Net.

Class	Class Name	Precision	Recall	Dice	IoU	HD95 ↓
0	Black Background	0.999	0.996	0.997	0.995	0.041
1	Abdominal Wall	0.993	0.992	0.992	0.985	0.231
2	Liver	0.985	0.986	0.985	0.971	0.116
3	Gastrointestinal Tract	0.971	0.970	0.970	0.942	0.594
4	Fat	0.983	0.985	0.984	0.969	0.383
5	Grasper	0.960	0.963	0.962	0.926	0.746
6	Connective Tissue	0.959	0.969	0.964	0.931	0.348
7	Blood	0.931	0.900	0.915	0.844	1.703
8	Cystic Duct	0.947	0.630	0.756	0.608	0.806
9	L-hook Electrocautery	0.970	0.978	0.974	0.950	0.169
10	Gallbladder	0.973	0.980	0.976	0.953	0.379
11	Hepatic Vein	0.907	0.919	0.913	0.839	0.599
12	Liver Ligament	0.989	0.993	0.991	0.982	0.000

model results remain close to the large structures with a mean Precision of 0.965, Recall of 0.970, Dice of 0.968, IoU of 0.937 and HD95 of 0.464. These structures are smaller and change their shape and position between frames. Due to these boundaries are more difficult to localise. The attention gates on the skip connections tackle these problems by suppressing the surrounding background before the fusion, maintaining a strong result. Among the minor classes, which include Hepatic Vein, Blood, Cystic Duct and Liver Ligament, the proposed model obtains a mean Precision of 0.944, Recall of 0.860, Dice of 0.894, IoU of 0.818 and HD95 of 0.777. These classes have very few pixels and thin shapes, and the transformer blocks applied at the deeper skip connections provide the surrounding context which is required to separate them from the neighbouring tissue of similar appearance. In the group of minor classes, the Cystic Duct is the only exception in performance, which has a slightly low recall of 0.630. However this class has a precision of 0.947, which indicates the predicted pixels are almost always correct, but some part of the class is left unsegmented; this is expected as it occupies only 0.04% of the annotated pixels. An important observation is that the precision and the recall remain close to each other for all classes except Cystic Duct. This balance indicates that the model neither over-segments nor under-segments these structures, and it is the reason for the consistently high Dice and IoU values. The HD95 values also remain low for most of the classes, which indicates accurate boundary alignment. The higher value obtained for Blood (1.703) is expected because of its irregular appearance, scattered distribution and annotation ambiguity. To examine whether this performance is better than the existing methods, the proposed model is compared with six state-of-the-art segmentation architectures in the next section.

4.4 Comparative Evaluation of the Proposed Model:

The results presented so far establish the performance of the proposed model, but they do not indicate whether this performance is better than that of the existing methods. Now, we will assess the proposed model's performance in comparison with the performance of six existing models: U-Net, Attention U-Net, ResUNet, Unet++, Swin-UNet, and TransUNet. The comparison has been carried out in two ways as some of the anatomical structures in the CholecSeg8k dataset contain much smaller number of pixels, i.e. Blood, Liver Ligament, Cystic Duct, and Hepatic Vein, as compared to other anatomical structures in the dataset: (i) overall mean performance across 12 anatomical classes excluding background class, and (ii) focused evaluation on clinically challenging minor classes which has less pixel representations (Blood, Liver Ligament, Cystic Duct, and Hepatic Vein).

4.4.1 Overall Mean Performance Analysis:

The mean results of all the models on 12 classes (excluding the background class) are presented in Table 4. The ResUNet has the comparable result due to large receptive field by stacking residual units throughout the network; however, it requires a high computational cost. The Swin-Unet Model gives a Dice of 0.944 and an IoU of 0.900 which are close to the proposed model but it restricts its attention within local windows and does not retain the convolutional inductive bias. The TransUNet model uses full

self-attention, whose cost is quadratic in the number of tokens and it reaches a Dice of 0.924 and an IoU of 0.869 which are much less as compared to that of the proposed model. Attention U-Net filters the skip features spatially but derives its gating signal only from the coarser decoder features and cannot introduce global context, obtaining a Dice of 0.935 and IoU of 0.887. The UNet++ narrows the semantic gap through nested skip pathways, but these remain purely local transformations, provides a competitive result on dice and IoU but poor results on HD95. The plain U-Net concatenates the encoder features into the decoder without any filtering and has a limited receptive field, and it obtains the lowest performance. The proposed model achieves the best results among the models used for comparison with a Dice of 0.949 and an IoU of 0.908, indicating that the employment of transformers to refine deep skip connections and hybrid loss is beneficial for the segmentation of organs. The proposed model benefits from both paradigms: the convolutional encoder provides strong local priors, while the lightweight transformer refines only the deep, low-resolution skip connections and thereby adds global context at a limited cost.

Table 4: Comparison with state-of-the-art models. Mean over 12 classes, excluding the background class.

Model	Precision	Recall	Dice	IoU	HD95 ↓
Hybrid U-Net (Proposed)	**0.964**	**0.939**	**0.949**	**0.908**	**0.506**
ResUNet	0.955	0.926	0.938	0.891	0.820
Swin-UNet	0.957	0.935	0.944	0.900	0.625
Attention U-Net	0.950	0.938	0.942	0.898	0.567
UNet++	0.954	0.927	0.938	0.891	0.745
TransUNet	0.952	0.925	0.936	0.888	0.710
U-Net	0.924	0.909	0.914	0.852	1.403

4.4.2 Performance Analysis on Minor classes:

Small anatomical structures, identified from the low pixel representation in Table 1, Hepatic Vein (16,478), Cystic Duct (62,178), Liver Ligament (732,447) and Blood (742,081) present significant segmentation challenges due to severe class imbalance and boundary ambiguity. These are also the organs where a wrong prediction carries the most risk during surgery, since injury to the cystic duct or the hepatic vein is one of the most serious complications in laparoscopic cholecystectomy. As shown in Table 5, ResUNet is the strongest model other than the proposed model on this group, with an IoU of 0.806 and a Dice of 0.886, because its residual connections allow it to build a deeper network without losing gradient strength, which helps it recover somewhat from the very low pixel count of these classes. Its HD95 is 1.032 because the residual blocks are still local convolutional operations and do not bring in information from distant parts of the image, so the boundary of a very thin structure like the cystic duct is still hard to place correctly. Swin-UNet reaches an IoU of 0.800 with the highest precision in the comparison after the proposed model (0.927), because its window-based attention lets it look at a wider area than a plain convolution before deciding

a pixel's class, which helps it avoid false positives on the rare classes. Its recall is comparatively low (0.853), because the attention is computed only within fixed local windows, and the interaction between windows is limited, so it can miss the parts of a thin structure that fall across a window boundary. Attention U-Net also reaches an IoU of 0.800, with the highest recall among other models (0.869), because its attention gates are trained to focus on the foreground and therefore recover more of these small structures than the plain convolutional models. Its precision is lower (0.906), because the gating signal comes only from the coarser decoder feature and does not carry any information about the wider context of the image, so it tends to include some background pixels along with the structure it is trying to recover. UNet++ obtains an IoU of 0.778, lower than ResUNet and the two models above. Its nested skip pathways are still built entirely from local convolutions, so they narrow the gap between the encoder and the decoder features without adding any information from outside the receptive field of the convolution, which limits its capability on the small structures. TransUNet obtains an IoU of 0.773 and the second-highest HD95 in the comparison (1.118). Its self-attention is applied only in the bottleneck, at the lowest resolution of the network, so the high-resolution skip features through which the fine boundaries of these structures are normally recovered are never refined, and the boundary accuracy on these classes suffers as a result. Plain U-Net gives the weakest result on every measure (IoU 0.728, HD95 2.272). Its skip connections transfer the encoder features to the decoder without any filtering, and its receptive field grows only gradually with depth, so it has neither the contextual information nor the selective filtering that the small, low-contrast structures in this group require. The proposed model obtains the best result on all five measures (Precision 0.944, Recall 0.860, Dice 0.894, IoU 0.818, HD95 0.777) because it combines the two properties that the baselines above have only one at a time. The transformer applied at the two deepest skip connections gives every position in the feature map information from the rest of the image, in the same way as Swin-UNet and TransUNet. This allows a very small and faint structure to be recognised using its surrounding context rather than being lost during pooling. The attention gate that follows then removes the background information this introduces, in the same way as Attention U-Net, which is what keeps the precision high. Because the two operations are applied one after the other on the same skip connections, the proposed model does not have to trade recall for precision as the other architectures do, and this is why its advantage over the baselines is largest on this smaller and riskier group of organs.

Table 5: Performance comparison on the minor classes

Model	Precision	Recall	Dice	IoU	HD95 ↓
Hybrid U-Net (Proposed)	**0.944**	**0.860**	**0.894**	**0.818**	**0.777**
ResUNet	0.927	0.855	0.886	0.806	1.032
Swin-UNet	0.927	0.853	0.882	0.800	1.027
Attention U-Net	0.906	0.869	0.882	0.800	0.824
UNet++	0.920	0.834	0.868	0.778	1.213
TransUNet	0.914	0.830	0.865	0.773	1.118
U-Net	0.868	0.813	0.834	0.728	2.272

4.5 Statistical Stability Analysis (Mean ± Standard Deviation)

In addition to the organ-wise performance analysis, the stability of this model is further assessed using per-image mean and standard deviation (SD) metrics on the test dataset, as summarised in Table 6. This table presents presents precision, recall, Dice, and IoU independently for each test image and reports the mean and standard deviation of these per-image values, with images in which a class is entirely absent excluded from that class's statistics. This analysis provides deeper insight into the consistency of the model's predictions in various samples. The findings show that the model retains high stability for major anatomical structures, with very low SD values observed for classes such as liver (±0.011), connective tissue (±0.013), and liver ligament (±0.002). These low variations suggest that the model produces reliable and consistent segmentation for well-defined and frequently occurring organs, complementing the strong performance observed in the previous subsection. For moderately complex structures such as grasper (±0.054) and gallbladder (±0.095), slightly higher variability is observed, which can be attributed to shape variations, occlusions, and surgical tool interactions. Furthermore, classes such as gastrointestinal tract (±0.112) and cystic duct (±0.202) exhibit relatively high standard deviation, reflecting greater prediction variability due to structural complexity and limited representation in the dataset. Despite these challenges, the model achieves competitive mean Dice scores for minor classes, including hepatic vein (0.934 ± 0.058) and blood (0.884 ± 0.075), demonstrating its robustness under class imbalance conditions. Overall, the low standard deviation across most classes confirms that the proposed model is not only accurate but also stable and generalizable across diverse samples.

Table 6: Per-class performance (Mean ± Standard Deviation) of the proposed Hybrid U-Net model

Class Name	Precision	Recall	Dice	IoU
Black Background	0.996 ± 0.037	0.995 ± 0.008	0.995 ± 0.032	0.991 ± 0.037
Abdominal Wall	0.985 ± 0.055	0.984 ± 0.040	0.983 ± 0.060	0.970 ± 0.068
Liver	0.981 ± 0.012	0.983 ± 0.015	0.982 ± 0.011	0.964 ± 0.021
Gastrointestinal Tract	0.950 ± 0.101	0.944 ± 0.084	0.939 ± 0.112	0.898 ± 0.124
Fat	0.980 ± 0.048	0.981 ± 0.020	0.979 ± 0.044	0.962 ± 0.053
Grasper	0.948 ± 0.055	0.954 ± 0.048	0.950 ± 0.054	0.908 ± 0.071
Connective Tissue	0.958 ± 0.017	0.969 ± 0.017	0.963 ± 0.013	0.930 ± 0.024
Blood	0.896 ± 0.094	0.878 ± 0.076	0.884 ± 0.075	0.799 ± 0.105
Cystic Duct	0.940 ± 0.021	0.907 ± 0.205	0.901 ± 0.202	0.855 ± 0.193
L-hook Electrocautery	0.961 ± 0.094	0.979 ± 0.021	0.966 ± 0.093	0.942 ± 0.094
Gallbladder	0.949 ± 0.097	0.965 ± 0.048	0.953 ± 0.095	0.919 ± 0.106
Hepatic Vein	0.923 ± 0.071	0.949 ± 0.074	0.934 ± 0.058	0.881 ± 0.099
Liver Ligament	0.990 ± 0.005	0.994 ± 0.004	0.992 ± 0.002	0.984 ± 0.005

4.6 Qualitative Evaluation

The qualitative results presented in Fig. 7 illustrate a visual assessment of the proposed model's segmentation capability. Each row contains the original input image,

its corresponding ground-truth annotation, and the segmentation output generated by the proposed method. As observed from the visual comparisons, the predicted regions closely match with the ground truth annotations, demonstrating accurate region localization and consistent boundary delineation across multiple anatomical structures. The model demonstrates an ability to preserve broad regions while recovering subtle structural details. The predicted segmentation maintains high alignment with ground truth

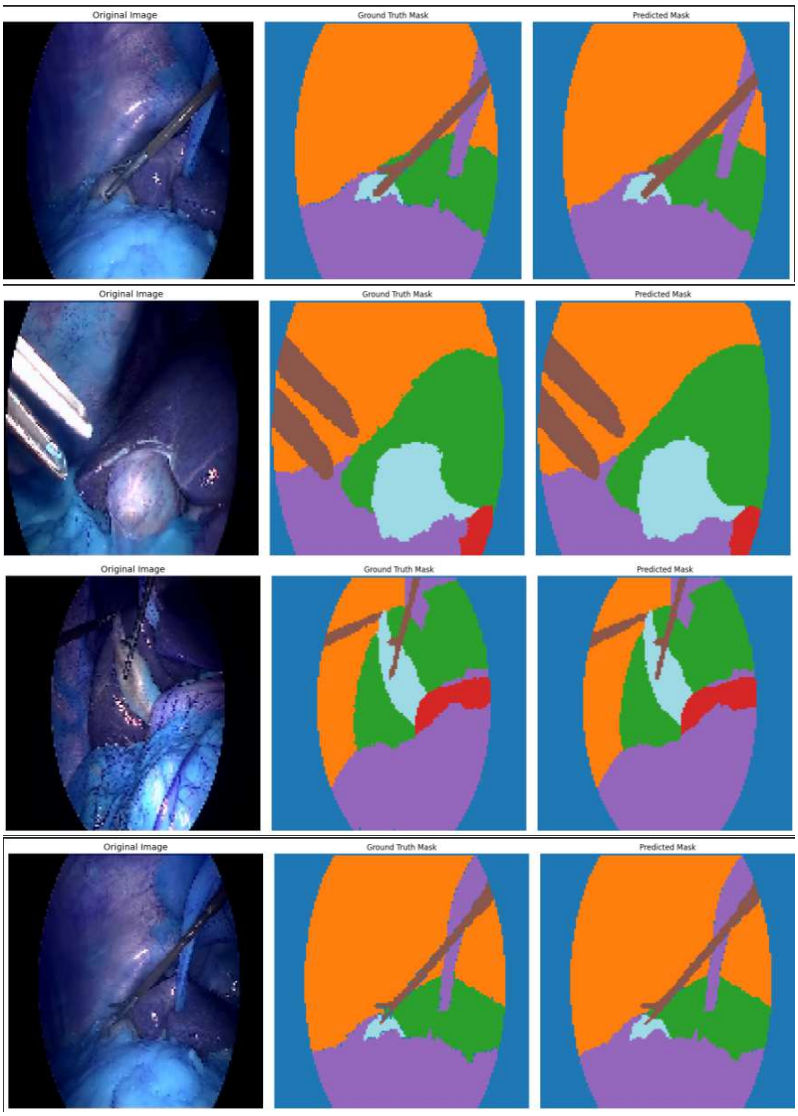


Fig. 7: Qualitative results of the proposed model. From left to right: input image, ground truth mask, and predicted mask.

boundaries, reflecting the effectiveness of the hybrid architecture in preserving both global contextual information and local spatial features. These qualitative results are consistent with the quantitative findings, confirming that the model produces accurate and spatially consistent predictions for anatomically small and challenging regions as well as for large and well-defined structures.

4.7 Ablation Study of the Proposed Loss Function

To verify the effectiveness of the proposed hybrid loss, we train the proposed architecture three times by keeping the architecture, dataset split, augmentation and all training hyperparameters fixed, and changing only the loss function. We compare the proposed hybrid loss ($\lambda = 0.3$) with its two individual components, i.e., Sparse Categorical Cross-Entropy ($\lambda = 1$) and Lovász-Softmax ($\lambda = 0$). Mean values of metrics are computed over 12 classes by excluding the background class. The results are given in Table 7. As evident from this Table 7, the proposed model with our new hybrid

Table 7: Ablation study of the loss function.

	All classes (excluding background)				Minor classes	
Loss Function	Precision	Recall	Dice	IoU	Dice	IoU
Cross-Entropy ($\lambda = 1$)	0.946	0.925	0.933	0.884	0.856	0.761
Lovász-Softmax ($\lambda = 0$)	0.947	**0.931**	0.936	0.888	0.867	0.776
Proposed Hybrid ($\lambda = 0.3$)	**0.955**	0.927	**0.938**	**0.891**	**0.869**	**0.780**

loss function gives the best result in terms of Precision, Dice and IoU, both over all the classes and on the minor classes. It achieves a mean Dice of 0.938 and IoU of 0.891, compared to 0.933 and 0.884 for using Cross-Entropy loss and 0.936 and 0.888 for Lovász-Softmax loss. The important observation here is that the hybrid loss performs better than both of its own components used separately, which shows that the two terms are not redundant and the combination is actually required. Results confirm that the proposed hybrid model effectively addresses pixel imbalance and fine-structure segmentation, enabling more accurate and spatially consistent predictions for anatomically small and challenging regions.

The results presented in this section establish the performance of the proposed model. The class-wise analysis shows that the model gives consistent and accurate results across all the three groups of structures(the large, moderately complex classes and rare structures). The comparison with six state-of-the-art models shows that the proposed model attains the best result on every metric over all the twelve classes and over the four minor classes separately, and that its advantage is concentrated on the structures that are hardest to segment. The statistical analysis on the test dataset demonstrates the relatively low standard deviations observed for most classes on metrics: Precision, Recall, Dice and IoU. This indicate that the proposed model provides not only high segmentation accuracy but also reliable and consistent predictions, supporting its robustness and generalization capability. The ablation study confirms that

both components of the hybrid loss are necessary, since removing either one reduces the performance, and that the combination improves the minor classes the most, which is consistent with the motivation for including the Lovász-Softmax term. In summary, these results show that the improvement brought by the proposed model does not come from a general increase in model size or from better performance on the easy, large structures, but from the specific combination of global context and selective filtering applied at the skip connections, which is where the benefit of the proposed design is most clearly visible.

5 Conclusion

In this paper, we have discussed a hybrid U-Net architecture for multi-class abdominal organ segmentation. It integrates transformers at multiple levels to refine the features from the encoder for modeling contextual information before passing these features through attention-gated skip connections for improving the feature representation. To address severe class imbalance, a hybrid loss function combining Lovász-Softmax and sparse categorical cross-entropy has been employed. The proposed loss function also enables effective optimization of both overlap and boundary accuracy. Comparative evaluation against six state-of-the-art architectures, including transformer-based models and U-Net variants, the proposed model consistently outperforms the state-of-the arts models across multiple metrics on the publicly available CholecSeg8k Dataset. This model can be helpful for the medical fraternity in surgical scene understanding and minimally invasive surgery.

Acknowledgements

I sincerely thank the UGC for providing the Junior Research Fellowship for my PhD work.

Declarations

- *The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.*

- Competing interests: *The authors have no relevant financial or non-financial interests to disclose*
- Ethics approval and consent to participate: Consent is not required as data is available in the public domain, and information is anonymized.
- Consent for publication:
- Data availability: Dataset used in this work is publicly available on the Kaggle Platform.
- Materials availability
- Code availability:
- Author contribution: All authors contributed equally in this work.

References

[1] Peng, K., Chang, Y., Lang, G., Xu, J., Gao, Y., Yin, J., Zhao, J.: Image segmentation network for laparoscopic surgery. Biomimetic Intelligence and Robotics **5**(3), 100236 (2025)

[2] Hong, W.-Y., Kao, C.-L., Kuo, Y.-H., Wang, J.-R., Chang, W.-L., Shih, C.-S.: Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. CoRR **abs/2012.12453** (2020)

[3] Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., *et al.*: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)

[4] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging **36**(1), 86–97 (2017)

[5] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)

[6] Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(6), 641–647 (1994)

[7] Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **8**(6), 679–698 (1986)

[8] Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision **1**(4), 321–331 (1988)

[9] Criminisi, A., Shotton, J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer, ??? (2013)

[10] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)

[11] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-24574-4_28 . (available also as arXiv:1505.04597 [cs.CV])

[12] Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 6813–6821 (2022)

[13] 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, pp. 3–11 (2018). Springer

[14] Wang, H., Cao, P., Yang, J., Zaiane, O.: Narrowing the semantic gaps in u-net with learnable skip connections: The case of medical image segmentation. Neural Networks **178**, 106546 (2024)

[15] Luo, Z., Zhu, X., Zhang, L., Sun, B.: Rethinking u-net: Task-adaptive mixture of skip connections for enhanced medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 5874–5882 (2025)

[16] Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4413–4421 (2018)

[17] Xia, S., Krishnan, S.M., Tjoa, M.P., Goh, P.M.: A novel methodology for extracting colon's lumen from colonoscopic images. Journal of Systemics, Cybernetics and Informatics **1**(2), 7–12 (2003)

[18] Arnold, M., Ghosh, A., Ameling, S., Lacey, G.: Automatic segmentation and inpainting of specular highlights for endoscopic imaging. EURASIP Journal on Image and Video Processing **2010**, 1–12 (2010) https://doi.org/10.1155/2010/814319

[19] Jimenez-Carretero, D., Fernandez-de-Manuel, L., Pascau, J., Tellado, J.M., Ramon, E., Desco, M., Santos, A., Ledesma-Carbayo, M.J.: Optimal multiresolution 3d level-set method for liver segmentation incorporating local curvature constraints. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 3419–3422 (2011) https://doi.org/10.1109/iembs.2011.6090925

[20] Marcinczak, J.M., Grigat, R.-R.: Closed contour specular reflection segmentation in laparoscopic images. International Journal of Biomedical Imaging **2013**, 1–6 (2013) https://doi.org/10.1155/2013/593183

[21] Nosrati, M.S., Peyrat, J.-M., Abinahed, J., Al-Alao, O., Al-Ansari, A., Abugharbieh, R., Hamarneh, G.: Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors. Lecture Notes in Computer Science, 324–331 (2014) https://doi.org/10.1007/978-3-319-10470-6_41

[22] Moccia, S., Wirkert, S.J., Kenngott, H., Vemuri, A.S., Apitz, M., Mayer, B., De Momi, E., Mattos, L.S., Maier-Hein, L.: Uncertainty-aware organ classification for surgical data science applications in laparoscopy. IEEE Transactions on Biomedical Engineering **65**(11), 2649–2659 (2018) https://doi.org/10.1109/TBME.2018.2813017

[23] Nosrati, M.S., Peyrat, J.-M., Abinahed, J., Al-Alao, O., Al-Ansari, A., Abugharbieh, R., Hamarneh, G.: Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors. In: Goll, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, vol. 8674, pp. 324–331. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_41

[24] Kolbinger, F.R., Rinner, F.M., Jenke, A.C., Carstens, M., Krell, S., Leger, S., Distler, M., Weitz, J., Speidel, S., Bodenstedt, S.: Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise–an experimental study. International Journal of Surgery **109**(10), 2962–2974 (2023)

[25] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation, 234–241 (2015). Springer

[26] Lei, R., Yang, M.: 5 d u-net for abdominal multi-organ segmentation. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 76–83. Springer, ??? (2023)

[27] Scheikl, P.M., Laschewski, S., Kisilenko, A., Davitashvili, T., Müller, B., Capek, M., Müller-Stich, B.P., Wagner, M., Mathis-Ullrich, F.: Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. Current Directions in Biomedical Engineering **6**(1), 20200016 (2020) https://doi.org/10.1515/cdbme-2020-0016

[28] Ghobadighadikalaei, V., Ismail, L.I., Hasan, W.Z.W., Ahmad, H., Ramli, H.R., Norsahperi, N.M.H., Tharek, A., Hanapiah, F.A.: Use of deep learning for liver segmentation during laparoscopic cholecystectomy. In: 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC), pp. 82–86 (2024). https://doi.org/10.1109/ICSGRC62081.2024.10690868

[29] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. IEEE Transactions on Medical Imaging **39**(5), 1856–1867 (2019)

[30] 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. In: Medical Imaging with Deep Learning (2018)

[31] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. (2017)

[32] Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices, 6848–6856 (2018)

[33] Poudel, R.P., Liwicki, S., Cipolla, R.: Fast-scnn: Fast semantic segmentation

network. In: British Machine Vision Conference (2019)

[34] Steger, T., Kletz, S., Schoeffmann, K.: Lite-rassp: A lightweight and real-time semantic segmentation network for surgical instrument detection. Healthcare Technology Letters (2022)

[35] Khan, R.F., Lee, B.-D., Lee, M.S.: Transformers in medical image segmentation: a narrative review. Quantitative Imaging in Medicine and Surgery **13**(12), 8747 (2023)

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need, 5998–6008 (2017)

[37] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020)

[38] 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 preprint arXiv:2102.04306 (2021)

[39] Valanarasu, J.M.J., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation, 36–46 (2021). Springer

[40] Petit, O., Thome, N., Rambour, C., Themyr, L., Collins, T., Soler, L.: U-net transformer: Self and cross attention for medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging, pp. 267–276 (2021). Springer

[41] Dan, Y., Jin, W., Yue, X., Wang, Z.: Enhancing medical image segmentation with a multi-transformer u-net. PeerJ **12**, 17005 (2024)

[42] Peng, Y., Chen, D.Z., Sonka, M.: U-net v2: Rethinking the skip connections of u-net for medical image segmentation. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2025). https://doi.org/10.1109/ISBI60581.2025.10980742