RADIOMICS LUS TRANSFORMER FOR LIVER TUMOUR SEGMENTATION

Ali Khalife[1], Mohammad Alkhatib[1], Erol Ozgur[1], Emmanuel Buc[2], Bertrand Le Roy[3], Youcef Mezouar[1], Adrien Bartoli[2]

[1] Institut Pascal, Clermont Auvergne INP, Clermont-Ferrand, France
[2] University Hospital, Clermont-Ferrand, France
[3] University Hospital, Saint-Etienne, France

ABSTRACT

Accurately segmenting liver tumours in laparoscopic ultrasonography (LUS) is critical for computer-assisted intervention guidance with augmented reality (AR). However, LUS images are often degraded by speckle noise, low contrast, and acoustic artifacts, while the liver's complex anatomy and tumour heterogeneity further complicate segmentation. These challenges are exacerbated by the probe's limited field of view and restricted manoeuvrability, which frequently produce partial or inconsistent tumour appearances across frames. Moreover, isoechoic tumours exhibit echogenicity similar to the surrounding liver tissue, rendering them nearly invisible in individual frames. To address these limitations, we propose `RadioLUS`, a hybrid deep-learning framework that operates on sequences of LUS frames. It integrates hierarchical visual features from a Swin-UNet backbone with handcrafted radiomics descriptors that capture speckle, texture, and shape characteristics. By aggregating temporal information across frames, the framework enhances weak tumour boundaries and improves the detection of subtle tissue variations. Experimental results demonstrate that `RadioLUS` outperforms the state-of-the-art under challenging LUS imaging conditions.

***Index Terms*—** laparoscopic ultrasound, tumour, segmentation, transformer, speckle, texture, shape

1. INTRODUCTION

Precisely localising tumours is essential in laparoscopic and robot-assisted liver surgery. Laparoscopic ultrasonography (LUS), which provides real-time intraoperative imaging, is increasingly used in this respect [1]. LUS images may however be challenging to interpret and tumours difficult to keep track of. It is thus be desirable to develop specific computer-assisted intervention means [2]. The task at end is *automatic liver tumour segmentation in LUS*. It is highly challenging due to the combined effects of image degradation and anatomical variability: LUS images are affected by speckle noise, low contrast, and acoustic artifacts that obscure tumour boundaries [3], while the liver's complex structure and tumour heterogeneity in size, shape, and echogenicity further complicate consistent segmentation [4]. Additional difficulties arise from the LUS probe's limited field of view and restricted manoeuverability [5], which often cause partial visibility and inconsistent tumour appearance across frames. Isoechoic tumours pose an especially severe challenge: their echogenicity closely matches the parenchyma's, rendering them nearly invisible in individual frames. These tumours are only visible via subtle variations in texture and speckle patterns discernible across consecutive frames [6].

The state-of-the-art is achieved by deep learning. The current methods typically rely on single-frame analysis [7, 8, 9, 10] and struggle with frame-level noise, weak tumour boundaries, and limited contextual information. To overcome these limitations, we propose `RadioLUS`, a hybrid deep learning framework that operates on LUS image *sequences*. `RadioLUS` integrates hierarchical visual features from a Swin-UNet backbone with handcrafted radiomics descriptors capturing speckle, texture, and shape characteristics. Temporal aggregation across frames reinforces weak tumour boundaries and enhances the visibility of subtle tissue patterns, while radiomics descriptors contribute domain-specific cues that improve segmentation accuracy and robustness. `RadioLUS` achieves more reliable tumour segmentation compared to single-frame methods and a sequence-based Swin-UNet baseline.

2. METHODOLOGY

We give `RadioLUS`' architecture. `RadioLUS` receives a short temporal sequence of T images $\mathcal{X} \in \mathbb{R}^{B\times T\times H\times W}$ as input, where B is the batch size, and (H, W) are the image dimensions. The network has three main stages: *(i)* hierarchical feature encoding and decoding using a Swin-UNet backbone [10], *(ii)* ultrasound descriptor extraction producing spatio-temporal radiomics features [6], and *(iii)* multi-stage Feature-wise Linear Modulation (FiLM) [11] conditioning, injecting radiomics-guided priors into the feature hierarchy. The final decoder yields a dense tumour probability map $\widehat{\mathbf{Y}} \in [0, 1]^{B\times H\times W}$.

2.1. Hierarchical Visual Feature Extraction

The visual backbone follows a 3D extension of Swin-UNet [10], combining windowed self-attention over space–time with U-shaped skip connections. Let $\mathbf{X}_{b,t} \in \mathbb{R}^{H\times W}$ denote the t-th grayscale ultrasound frame of sequence b, with $t \in \{1,\dots,T\}$. Each frame is divided into non-overlapping patches of size $s \times s$, which are linearly projected into D-dimensional token embeddings:

$$\mathbf{z}_{b,i,p} = \mathbf{W}_{\text{proj}}\,\text{vec}\big(\mathcal{P}_{b,i,p}\big) + \mathbf{p}_{i,p}, \tag{1}$$

where $\mathcal{P}_{b,i,p}$ denotes the p-th patch in tubelet i, $\mathbf{p}_{i,p}$ is a learnable spatio-temporal positional embedding, and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{D\times(s^2\tau)}$ is the patch projection matrix. Here, D is the embedding dimension and τ is the temporal tubelet length, yielding $T_v = T/\tau$ tubelets per sequence.

The patch tokens $\{\mathbf{z}_{b,i,p}\}$ extracted from each tubelet are reshaped into a spatial–temporal tensor $\mathbf{F}$, serving as the input to the hierarchical Swin-UNet encoder. The encoder is composed of L stages indexed by ℓ, each performing shifted-window attention and patch merging to progressively model context and build multi-scale representations:

$$\mathbf{F}_\ell = \mathcal{E}_\theta^{(\ell)}(\mathbf{F}_{\ell-1}), \qquad \ell = 1,\dots,L, \tag{2}$$

where $\mathbf{F}_\ell \in \mathbb{R}^{B\times T_v\times H_\ell\times W_\ell\times D_\ell}$ denotes the stage-ℓ feature tensor. Skip connections forward encoder features to the decoder to preserve boundary details and fine spatial structures.

2.2. Spatio-temporal Radiomics Descriptors Extraction

To capture the effects of texture, speckle and shape over time, we extract spatio-temporal radiomics descriptors aligned to the Swin token grid. For each tubelet (b,i,p), a C-dimensional descriptor vector $\mathbf{s}_{b,i,p} \in \mathbb{R}^C$ is computed by aggregating three complementary feature families.

(i) Texture. We apply Mean Subtracted Contrast Normalization to stabilize local contrast, followed by a neural Local Binary Pattern operator that computes differentiable, learnable binary patterns [12, 13]. This combination captures fine-grained texture and intensity co-occurrences while remaining robust to gain and contrast variations.

(ii) Speckle. Local scattering behaviour is quantified through envelope statistics. Within each neighborhood, the local mean and variance are computed to characterize the speckle contrast, with a small term added to ensure numerical stability. These local statistics are also used to estimate the parameters of a Nakagami distribution [14], which models the echo envelope. Here, the shape and scale parameters are derived from the moments of the envelope, linking the local mean and variance to the distribution's spread and average signal power.

(iii) Shape. Geometric descriptors such as compactness, eccentricity, and orientation are predicted from intermediate segmentation maps [6]. Auxiliary representations (boundary, signed distance, and skeleton maps) reinforce contour smoothness, while global Zernike and Fourier descriptors encode overall morphology.

Finally, the radiomics features are combined, by stacking all descriptors to form the tensor $\mathcal{S} = \{\mathbf{s}_{b,i,p}\} \in \mathbb{R}^{B\times T_v\times P\times C}$. Each descriptor vector is channel-normalized per sequence and linearly projected into the model's embedding space through a learnable transformation:

$$\mathbf{r}_{b,i,p} = \mathbf{W}_{\text{rad}}\,\mathbf{s}_{b,i,p} + \mathbf{b}_{\text{rad}}, \qquad \mathbf{W}_{\text{rad}} \in \mathbb{R}^{D\times C}. \tag{3}$$

This produces radiomics tokens $\mathbf{r}_{b,i,p} \in \mathbb{R}^D$ spatially aligned with the initial Swin-UNet token grid $\mathbf{F}_\ell$. These features are later integrated into the encoder–decoder hierarchy via FiLM conditioning, allowing physics-informed modulation of the transformer activations.

2.3. Multi-stage FiLM Conditioning

To fuse radiomics-informed cues with learnt visual representations, we apply FiLM [11] at each encoder-decoder stage. At level ℓ, the projected radiomics tokens $\mathbf{r}_{b,i,p}$ are first processed by a small multilayer perceptron to generate modulation parameters:

$$[\boldsymbol{\gamma}^\ell_{b,i,p},\, \boldsymbol{\beta}^\ell_{b,i,p}] = \text{MLP}_\ell(\mathbf{r}_{b,i,p}), \tag{4}$$

where $\boldsymbol{\gamma}^\ell_{b,i,p}, \boldsymbol{\beta}^\ell_{b,i,p} \in \mathbb{R}^{D_\ell}$ define per-channel scaling and shifting coefficients. These parameters are applied to the corresponding Swin feature maps by an affine transformation:

$$\widehat{\mathbf{F}}_\ell(b,i,p) = \boldsymbol{\gamma}^\ell_{b,i,p} \odot \mathbf{F}_\ell(b,i,p) + \boldsymbol{\beta}^\ell_{b,i,p}, \tag{5}$$

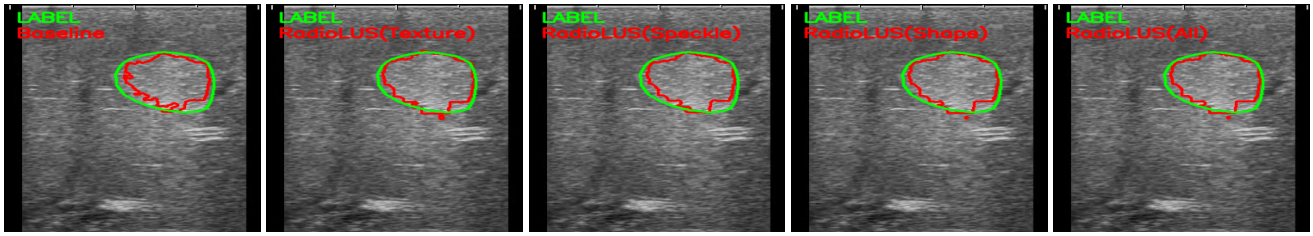
where $\odot$ denotes element-wise multiplication. Multi-scale conditioning allows coarse radiomics priors (*e.g.*, global shape and intensity statistics) to guide deeper semantic layers, while fine-grained texture and speckle cues modulate early representations. This stage-wise fusion aligns deep activations with ultrasound physics, improving boundary precision and robustness under variable imaging conditions.

2.4. Decoder and Loss Formulation

The modulated feature hierarchy $\{\widehat{\mathbf{F}}_\ell\}$ serves as input to the Swin-UNet decoder, which progressively upsamples and fuses skip-connected encoder features to recover fine spatial details [10]. At each stage, concatenated encoder and decoder activations undergo windowed self-attention and patch expansion, yielding refined representations that integrate global context with boundary-level precision. The final decoder layer outputs a tumour probability map for the central frame of each input sequence, $\widehat{\mathbf{Y}}_b \in [0,1]^{H\times W}$. The model parameters are optimised using a hybrid objective:

$$\mathcal{L} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}(\widehat{\mathbf{Y}},\mathbf{Y}) + \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}}(\widehat{\mathbf{Y}},\mathbf{Y}), \tag{6}$$

where $\mathbf{Y}$ denotes the ground-truth segmentation mask. The Dice term promotes overlap and the Binary Cross-Entropy (BCE) term stabilizes training under class imbalance.



IoU 66.79%, Dice 80.09% IoU 83.83%, Dice 91.20% IoU 78.04%, Dice 87.66% IoU 81.41%, Dice 89.75% IoU 84.07%, Dice 92.15%

Fig. 1: Example results from ablation study. Green contour is the manual label, and red is the `RadioLUS` prediction.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Dataset and Setup

We collected 17 intraoperative LUS videos from 17 laparoscopic liver resections; data collection is IRB approved. All data were obtained with informed patient consent and approval, and were fully anonymized prior to analysis. Videos were recorded at 39 fps and temporally subsampled to 10 fps ($\Delta t = 0.1$ s). Each training sample consists of a sequence of four consecutive frames ($T = 4$) representing short-term temporal evolution of the scene. A total of 11853 frames were manually segmented by expert surgeons and verified by consensus. The data were grouped into 14 training videos (9332 frames), 1 validation video (1052 frames), and 2 test videos (1469 frames). The frames were resized to 224×224 and normalised using training-set statistics. Temporal sequences were generated online during loading to preserve frame continuity. We applied data augmentations using Albumentations, including affine rotations ($\pm 10°$), translations ($\leq 5\%$), elastic and grid distortions, horizontal flips, and photometric adjustments (brightness and contrast). All experiments were implemented in PyTorch 2.0 with mixed-precision training and gradient clipping. Optimisation used AdamW with initial learning rate 5×10^{-4}, weight decay 1×10^{-3}, and cosine annealing scheduler. All FiLM MLPs and projection weights are trained end-to-end. The model was trained on an NVIDIA RTX A5000 (24 GB). Inference runs at ∼30 fps for 224×224 inputs, enabling real-time deployment in surgical workflows. We used $B = 8$, $s = 16$, $\tau = 2$, base embedding $D = 96$, $L = 4$ stages, $C = 8$, $\lambda_{\text{Dice}} = 0.6$ and $\lambda_{\text{BCE}} = 0.4$.

3.2. Results and Discussion

We quantify the robustness of tumour segmentation methods using 99% normal confidence intervals (CI) for the IoU and Dice scores. Using a test set of size $N = 1469$, the confidence intervals were estimated as $\mu \pm \epsilon_{99\%CI} = \mu \pm 2.576 \times \frac{\sigma}{\sqrt{N}}$, where μ and σ denote the mean and standard deviation of the per-image segmentation scores, respectively. These intervals provide an estimate of the uncertainty associated with the scores.

In terms of the baselines, we observe in table 1 that the single-frame Swin-UNet outperforms all other single-frame methods, while the multi-frame Swin-UNet globally outperforms. We thus consider the multi-frame Swin-UNet without radiomics descriptors as the state-of-the-art baseline. Its results are shown in the first row of the ablation result table 2.

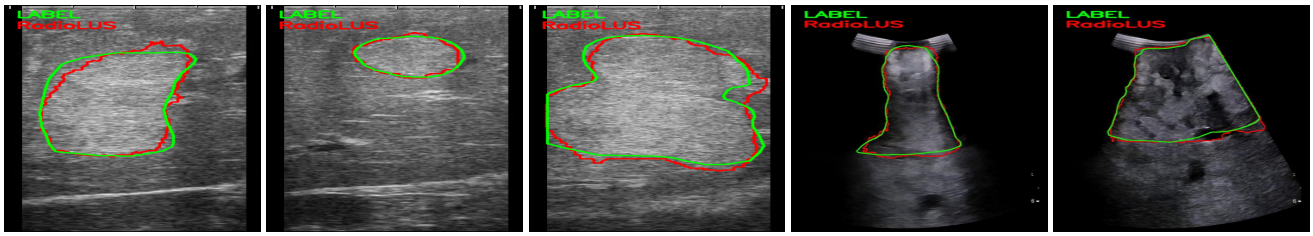
Table 1: Baseline methods ($\mu \pm \epsilon_{99\%CI}$).

Model	Frames	IoU %	Dice %
U-Net [7]	Single	84.05 ± 1.87	88.44 ± 1.75
FCN [8]	Single	84.52 ± 1.83	88.05 ± 1.72
TransUNet [9]	Single	86.32 ± 1.77	89.86 ± 1.74
Swin-UNet [10]	Single	86.00 ± 1.57	90.03 ± 1.47
Swin-UNet	Multi	**86.33 ± 1.48**	**90.89 ± 1.37**

Table 2: Ablation study for `RadioLUS` ($\mu \pm \epsilon_{99\%CI}$).

Texture	Speckle	Shape	IoU %	Dice %
–	–	–	86.33 ± 1.48	90.89 ± 1.37
✓	–	–	88.38 ± 1.49	92.59 ± 1.42
–	✓	–	87.24 ± 1.28	91.68 ± 1.10
–	–	✓	87.65 ± 1.70	91.92 ± 1.64
✓	✓	–	88.82 ± 1.24	92.94 ± 1.07
✓	–	✓	89.17 ± 1.40	93.24 ± 1.33
–	✓	✓	88.51 ± 1.29	92.73 ± 1.14
✓	✓	✓	**89.68 ± 1.18**	**93.55 ± 1.02**

The ablation results quantify how each radiomics parameter family contributes and how they interact when fused with the neural backbone. The visual-only baseline (Swin-UNet) already performs well with 86.33% IoU and 90.89% Dice, but each radiomics family is shown to bring an improvement. Figure 1 shows an example how each radiomics feature contributes. Texture descriptors give the largest single-family gain of 1.7 pp, indicating their strong role in sharpening local contrast and boundary cues. Shape descriptors alone also provide a substantial boost of 1 pp, reflecting the benefit of geometric regularisation for boundary fidelity. Speckle statistics yield smaller single-family gain of 0.8 pp but clearly complement the other features. Pairwise combinations further improve performance, with shape and texture achieving 2.35 pp gain. Finally, fusing all three families yields the largest im-



IoU 87.36%, Dice 93.25% IoU 88.90%, Dice 94.12% IoU 89.19%, Dice 94.29% IoU 90.91%, Dice 95.24% IoU 93.61%, Dice 96.70%

Fig. 2: Segmentation examples of `RadioLUS`. Green contour is the manual label, and red is the `RadioLUS` prediction.

provement of 2.66 pp. The pattern is clear: texture and shape provide the largest, immediately useful signals for boundary and region accuracy, while speckle adds complementary microstructural evidence that produces modest but consistent additional improvements when integrated via our multi-stage FiLM conditioning. The experiments show that our multi-stage FiLM fusion provides a compact, effective way to inject ultrasound radiomics into transformer backbones, yielding a good performance for intraoperative tumour segmentation. Figure 2 shows qualitative results for tumour segmentation in LUS images.

4. CONCLUSION

We have presented `RadioLUS`, a radiomics-guided video transformer for liver tumour segmentation in LUS images. It integrates speckle, texture, and shape descriptors via FiLM modulation, and achieves accuracy scores of Dice 93.55% and IoU 89.68% with real-time inference (30 fps). Future work will *(i)* train neural methods to achieve radiomics features extraction with additional invariance and covariance terms, and *(ii)* generalise to other organ tumours.

5. ACKNOWLEDGEMENTS

This work is funded by project ANR JCJC - IMMORTALLS.

6. COMPLIANCE WITH ETHICAL STANDARDS

This study used fully anonymised, retrospectively collected LUS images with IRB00008526-2019-CE58 approval and a waiver of informed consent.

7. REFERENCES

[1] Russolillo et al., "Ultrasound liver map technique for laparoscopic liver resections: tips and tricks," *Mini-invasive Surgery*, vol. 7, no. 3, 2023.

[2] Brockmeyer et al., "The role of augmented reality in the advancement of minimally invasive surgery procedures: A scoping review," *Bioengineering*, vol. 10, 2023.

[3] Almajalid et al., "Development of a deep-learning-based method for breast ultrasound image segmentation," *ICMLA*, 2018.

[4] Ryu et al., "Joint segmentation and classification of hepatic lesions in ultrasound images using deep learning," *European radiology*, vol. 31, no. 11, 2021.

[5] Schneider et al., "Performance of image guided navigation in laparoscopic liver surgery–a systematic review," *Surgical Oncology*, vol. 38, 2021.

[6] Gong et al., "Ultrasound image texture feature learning-based breast cancer benign and malignant classification," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[7] Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. 2015, pp. 234–241, Springer.

[8] Long et al., "Fully convolutional networks for semantic segmentation," *CVPR*, 2015.

[9] Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv*, 2021.

[10] Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv*, 2021.

[11] Pérez et al., "Film: Visual reasoning with a general conditioning layer," in *ICLR*, 2018.

[12] Xiao et al., "Blind quality metric via measurement of contrast, texture, and colour in night-time scenario," *KSII Trans. Internet Inf. Syst.*, 2021.

[13] Chang et al., "Multi-scale lbp fusion with the contours from deep cellnns for texture classification," *Expert Systems with Applications*, vol. 238, pp. 122100, 2024.

[14] Destrempes et al., "A critical review and uniformized representation of statistical distributions modeling the ultrasound echo envelope," *Ultrasound in medicine & biology*, vol. 36, no. 7, pp. 1037–1051, 2010.