Proceedings Paper:
Battye, J., Relton, S., Conaghan, P. et al. (Accepted: 2026) BEA-Net: Boundary-Aware 3D Attention Network for MRI Knee Cartilage Segmentation. In: IEEE Open Journal of Engineering in Medicine and Biology. The 39th IEEE International Symposium on Computer-Based Medical Systems, 03-05 Jun 2026, Limassol, Cyprus. IEEE. EISSN: 2644-1276. (In Press)

BEA-Net: Boundary-Aware 3D Attention Network for MRI Knee Cartilage Segmentation

James Battye*, Samuel Relton*, Philip Conaghan†, Ping Lu*
*University of Leeds, Leeds, UK
†NIHR Leeds Biomedical Research Centre, Leeds, UK

***Abstract*—Segmenting knee cartilage from magnetic resonance images is vital to understanding the pathogenesis and progression of knee osteoarthritis. However, it presents several challenges due to the complex morphology and thin structure of knee cartilage. Dedicated boundary learning has been shown to improve segmentation predictions from deep learning models when segmenting tissues with complex or unclear boundaries, but the application of dedicated boundary learning for 3D segmentation of knee cartilage from MRI has been limited. In this work, we introduce BEA-Net, a multi-task boundary-aware attention network for 3D segmentation of knee cartilage from MRI. BEA-Net uses a dual-branch architecture with an auxiliary decoder dedicated to learning boundary information. A novel boundary-enhancement attention module is used to amplify and focus on salient boundary features to refine boundary predictions. Learnt boundary features are fused with primary-decoder features to enhance segmentation predictions, and the network is optimised using a combination loss that encourages inter-decoder consistency. BEA-Net was evaluated on an MRI dataset from the Osteoarthritis Initiative, outperforming other state-of-the-art models when segmenting four types of knee cartilage and achieving Dice scores of 89.52%, 88.46%, 85.87%, 87.11% when segmenting the femoral, tibial, and patellar cartilage, and the meniscus respectively.**

***Index Terms*—Knee cartilage segmentation, MRI, convolutional neural networks, multi-task learning, spatial attention**

I. Introduction

Osteoarthritis (OA) is a common joint condition in adults, characterised by cartilage loss, joint stiffness, and chronic pain. The disease affects around 595 million people globally, an increase of 132.2% since 1990 [1]. OA most commonly affects the knee and almost 30% of over-45s have radiographic evidence of knee OA [2]. Late-stage knee OA often leads to knee replacement, and more than 100,000 such procedures are performed in the UK every year [3].

As knee OA progresses, the morphology of knee joint structures changes; osteophytes develop and cartilage degenerates. Studying the morphology of these structures is vital to understand knee OA pathogenesis and progression. Magnetic resonance imaging (MRI) allows imaging of soft tissue, such as cartilage, making it an effective imaging modality for studying knee OA. To analyse knee morphology from MRI, key joint structures must be accurately segmented. Manual segmentation is laborious, time consuming, and difficult to reproduce [4]. Automating segmentation overcomes these challenges.

Knee cartilage has complex morphology and articular cartilage is thin and sheet-like making automatic segmentation challenging. As knee OA progresses and cartilage degenerates, its smooth surface becomes rough and fissures appear further increasing the complexity of its morphology. Automatic MRI segmentation of knee cartilage has evolved from region-growing, edge-based algorithms, and statistical shape modelling to current state-of-the-art (SOTA) methods using deep learning [5]. Models based on U-Net [6] have demonstrated SOTA performance in several image segmentation domains, including knee cartilage segmentation [7].

Multi-task learning (MTL), where a model is jointly optimised to perform primary and auxiliary tasks, supports robust and generalisable learning by allowing models to draw on salient features in auxiliary tasks while acting as regularisation. Multiple studies have shown the utility of MTL in medical image segmentation [8]. Boundary detection is a fundamental task in computer vision, and there are several operators for boundary enhancement [9]. Explicit boundary modelling has been recognised as important in medical image segmentation and has produced SOTA results in numerous domains [10]. However, there has been limited application of these techniques for 3D segmentation of knee cartilage from MRIs.

Motivated by the challenges of accurately delineating thin knee cartilage with complex boundaries, this work introduces BEA-Net; a multi-task network for segmenting knee cartilage from MRI. BEA-Net uses a dual-branch architecture with an auxiliary decoder to learn cartilage boundaries via a novel Boundary Enhancement Attention Module (BEAM). Learnt boundary features complement segmentation predictions, and a consistency loss term regularises the model while encouraging both decoders to learn similar macroscopic structure. BEAM amplifies and maintains salient edges using Sobel filters and spatial attention. BEA-Net was evaluated on a subset of MRIs from the Osteoarthritis Initiative (OAI), demonstrating improved performance over other SOTA models. The contributions of this work can be summarised as:

- Proposing BEA-Net, a novel segmentation model with dedicated boundary learning facilitated by a novel boundary enhancement attention module;
- Introducing BEAM, a novel module which amplifies and filters for salient boundary information using Sobel filters and spatial attention; and
- Evaluating BEA-Net's ability to segment four types of knee cartilage using an MRI dataset from the OAI.

II. Method

*1) **Model Architecture**:* An overview of the model architecture is shown in Figure 1. BEA-Net used a single encoder and

two decoders. Encoder, primary decoder, and auxiliary decoder feature maps are denoted f_e^i, f_{ds}^i, and f_{db}^i, respectively, where $1 \leq i \leq 6$ denotes the model layer. The encoder successively downsampled input MRIs. The primary decoder performed upsampling before outputting a segmentation mask, M, for four types of cartilage tissue: the femoral, tibial, and patellar cartilage, and meniscus. The auxiliary boundary decoder performed upsampling with attention-guided boundary enhancement before generating a boundary mask, B, for the four tissues. Before the final segmentation prediction, boundary features, f_{db}^1, were fused with the main decoder features, f_{ds}^1.

Encoder blocks contained three convolutional blocks. Each block performed 3D convolution, batch normalisation, and ReLU activation. In each layer, the first two blocks performed 3×3 convolution, and the final block performed downsampling using 3×3 convolution with stride of two.

The primary decoder predicted a segmentation mask, M. Segmentation decoder blocks performed upsampling using 2×2 transpose convolution. Upsampled features were concatenated with the corresponding encoder skip connection, before passing through two convolutional blocks and a convolutional block attention module [11]. A fusion block then combined the final segmentation features and boundary logits before outputting a segmentation prediction for four cartilage classes.

The auxiliary decoder predicted a multiclass cartilage boundary mask, B. In each layer, features were up-sampled using 2×2 transpose convolution, combined with encoder skip connections, and passed through a BEAM block where attention-guided boundary enhancement was performed.

*2) **Boundary Enhancement Attention Module (BEAM)**:* In each BEAM block, input features $f^i \in \mathbb{R}^{C\times H\times W\times D}$, where C, H, W, and D denote channels, height, width, and depth, respectively, were first passed through a 3×3 convolutional block halving channel depth to $C/2$. This can be formulated as $f'^i = Block^{3\times3}(f^i)$, where f'^i is the processed feature map and $Block^{3\times3}$ is a convolutional block consisting of 3×3 convolution, batch normalisation, and ReLU activation. Features were then projected via 1×1 convolution reducing channel depth to $C/4$ and passed through a convolutional block. This can be formulated as $f''^i = Block^{3\times3}(F_{conv}^{1\times1}(f'^i))$ where $F_{conv}^{1\times1}$ is 1×1 convolution. Boundary enhancement was then performed using a 3D Sobel filter. Boundary-enhanced features were normalised, passed through a ReLU activation, then projected to a channel depth of $C/2$ as follows: $f_{bdry}^i = F_{conv}^{1\times1}(ReLU(BN(Sobel(f''^i))))$, where $Sobel$ is a 3D Sobel filter, BN is batch normalisation, and f_{bdry}^i is the boundary-enhanced features. Features were then refined using a spatial attention mechanism. Average and max pooling were performed across channels and concatenated to produce f_{pool}^i as follows: $f_{pool}^i = [MaxPool(f_{bdry}^i), AvgPool(f_{bdry}^i)]$. To learn richer spatial representations, the pooled features were projected via 1×1 convolution to $f_p'^i \in \mathbb{R}^{8\times H\times W\times D}$. Attention features were then normalised, passed through a ReLU activation, and projected back to a single channel: $f_{pool}'^i = F_{conv}^{1\times1}(ReLU(IN(F_{conv}^{1\times1}(f_{pool}^i))))$, where IN represents instance normalisation. Attention features underwent sigmoid activation, σ, to produce a spatial boundary attention map: $a_{bdry}^i = \sigma(f_{pool}'^i)$. The attention map was used to weight the original boundary-enhanced features as a residual connection: $f_{bdry}'^i = a_{bdry}^i \odot f_{bdry}^i + f_{bdry}^i$ where $\odot$ is the Hadamard product. The spatially weighted boundary features were concatenated with f'^i, projected to halve channel depth to $C/2$, and passed through a convolutional block: $f_{out}^i = Block^{3\times3}(F_{conv}^{1\times1}([f'^i, f_{bdry}'^i]))$.

*3) **Loss Functions**:* Predicted segmentation masks, M, were compared to ground truth masks, M_{GT}, using an equal weighting of Dice loss and cross entropy loss: $\mathcal{L}_s = \mathcal{L}_{Dice}(M, M_{GT}) + \mathcal{L}_{CE}(M, M_{GT})$. Boundary predictions, B, were compared to boundary ground truths, B_{GT}, using cross entropy loss: $\mathcal{L}_b = \mathcal{L}_{CE}(B, B_{GT})$. A consistency loss encouraged both decoders to learn similar macroscopic structure and act as regularisation. Structural similarity index measure (SSIM) was used to assess consistency between decoders. A consistency loss was implemented as deep supervision between f_{ds}^1 and f_{db}^1 as $\mathcal{L}_c = 1 - SSIM(f_{ds}^1, f_{db}^1)$. The overall objective function was therefore $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_b + \mathcal{L}_c$.

TABLE I
DATASET SPLITS SUMMARY. *KLG*: KELLGREN-LAWRENCE GRADE

Dataset	Patients	Samples	Age (yrs)	KLG1	KLG2	KLG3	KLG4
Train	74	148	45-78	3%	34%	57%	6%
Test	14	28	49-78	0%	32%	57%	11%

III. EXPERIMENTS

*1) **Dataset**:* BEA-Net was evaluated on a subset of MRI data from the OAI ("OAI Imorphics dataset") [12]. The subset contained 176 scans from 88 patients. Each sample was $384 \times 384 \times 160$ voxels with resolution of $0.36mm \times 0.36mm \times 0.7mm$. Ground truth masks were generated by a single expert at Stryker Imorphics. Boundary ground truths were derived from segmentation ground truths using erosion and dilation operations. The scans were split into train and test sets of 148 scans and 28 scans respectively, as shown in Table I. Five-fold cross-validation was performed using the train set. Each patients' scans were assigned to a single fold to avoid data leakage. Test results were obtained by applying a single trained model to the held-out test set.

*2) **Implementation Details**:* Model training used two NVIDIA L40S GPUs and was implemented in PyTorch. Patches of size $128 \times 128 \times 128$ were sampled from MRI volumes. Patches were sampled with equal probability that the patch centre belonged to the background, femoral, tibial, patellar cartilage, or meniscus class. Patches were augmented using intensity scaling, intensity shifting, and random flips along all three axes. Cross-validation training was performed for 360 epochs using an Adam optimiser. Learning rate was scheduled to anneal from 1e-5 to 1e-3 over the first 10% epochs before decaying following a cosine trajectory to 1e-8. If training appeared unstable or noisy, the warm-up period was

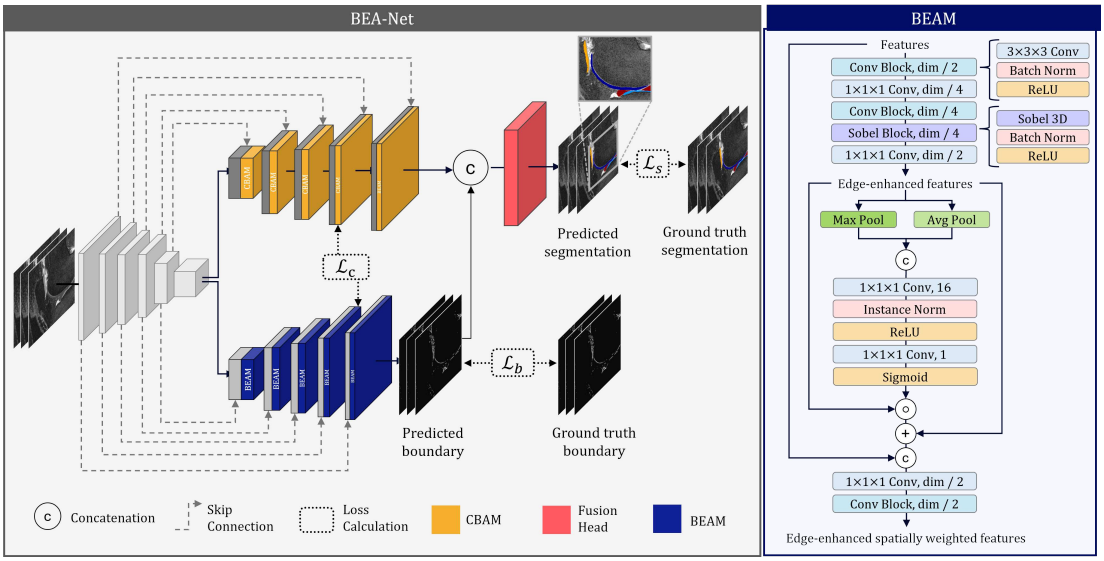

Fig. 1. *Left:* Architecture of BEA-Net which applies BEAM in the auxiliary decoder. *Right:* BEAM architecture

increased. Batch size was set to four. Predicted segmentation masks were assessed using Dice score (DSC), intersection over union (IOU), average symmetric surface distance (ASSD), Hausdorff distance (95th percentile [HD95]).

*3) **Baseline and Ablation Experiments**:* BEA-Net was assessed by segmenting the femoral, tibial, and patellar cartilage, and meniscus from the OAI Imorphics dataset. BEA-Net was compared to three other SOTA models: 3D U-Net [13], SwinUNETR-V2 [14], and Segformer3D [15]. SwinUNETR-V2 has achieved SOTA performance in several medical imaging domains, while Segformer3D achieved comparable results with fewer parameters. All models were implemented as described in the original papers. To investigate the effectiveness of BEAM components, an ablation study was conducted in which different versions of BEA-Net were trained using a single cross-validation fold for 500 epochs.

IV. Results

*1) **Baseline Comparison and Ablation Study**:* Average cross-validation results are shown in Table II. BEA-Net outperformed all other models for IOU and ASSD for most folds for all tissues. For IOU, on average, BEA-Net outperformed the next best model by 0.47%, 0.09%, 1.25%, and 0.76% for the femoral (FC), tibial (TC), and patellar cartilage (PC), and meniscus (M) respectively. For ASSD, on average, BEA-Net outperformed by 3.10%, 0.62%, 11.97%, and 3.56% for each tissue respectively.

Test set results are shown in Table III. BEA-Net outperformed all other models for most evaluation metrics in most tissues, achieving DSCs of 89.52%, 88.46%, 85.87%, and 87.11% for the femoral, tibial, and patellar cartilage, and meniscus respectively. BEA-Net achieved increases in IOU of 0.76%, 0.88%, 0.74%, and 0.61% for each tissue respectively compared to the next best model. These results emphasise the utility of BEA-Net's dedicated boundary learning branch and

TABLE II
Cross-validation results: fold-average IOU and ASSD.

	IOU (%) ↑				ASSD ↓			
	FC	TC	PC	M	FC	TC	PC	M
BEA-Net (Ours)	**81.09**	**79.97**	**72.79**	**76.79**	**0.237**	**0.238**	**0.457**	**0.359**
SwinUNETR-V2	80.70	79.90	71.90	76.12	0.244	0.239	0.519	0.373
U-Net	80.41	79.36	70.92	76.21	0.308	0.253	0.581	0.378
Segformer3D	66.43	67.82	54.10	62.46	0.603	0.480	1.302	0.885

BEAM while demonstrating its ability to perform robust knee cartilage MRI segmentation.

Ablation study results assessing the effectiveness of spatial attention within BEAM are shown in Table IV. "Edge" represents including edge enhancement via 3D Sobel filters and convolution, and "Attn" represents inclusion of spatial attention. Including spatial attention led to a performance uplift for most evaluation metrics. Surface-based metrics improved most for the tibial and patellar cartilage. For the tibial cartilage, ASSD reduced by 3.77% and HD95 reduced by 10.64%. For the patellar cartilage, ASSD reduced by 2.42% and HD95 reduced by 6.43%. This indicates spatial attention offers most utility for smaller tissues with complex morphology, such as the patellar cartilage.

*2) **Qualitative Results**:* Figure 2 presents predicted segmentation masks from all four models with the original images. These examples demonstrate BEA-Net's ability to segment small, thin cartilage structure accurately, such as the femoral cartilage in row one and two. BEA-Net's improved ability to segment several complex interfacing structures is shown in row three and four.

*3) **Complexity Analysis**:* To analyse computational complexity, the parameter count and inference speed of each model was assessed. Inference speed was measured by assessing the average time taken for each model to process 100 tensors of shape (1,1,128,128,128). SwinUNETR-V2 had the most parameters (62.19M), followed by BEA-Net (53.28M),

TABLE III
TEST SET RESULTS: AVERAGE IOU, DSC, ASSD AND HD95

	IOU (%) ↑				DSC (%) ↑				ASSD ↓				HD95 ↓			
	FC	TC	PC	M	FC	TC	PC	M	FC	TC	PC	M	FC	TC	PC	M
BEA-Net (Ours)	**81.08**	**79.40**	**75.82**	**77.26**	**89.52**	**88.46**	**85.87**	**87.11**	**0.221**	**0.275**	0.287	**0.352**	**0.930**	**1.961**	1.307	1.653
SwinUNETR-V2	80.33	78.71	75.26	76.79	89.06	88.01	85.68	86.80	0.239	0.281	**0.269**	0.359	0.972	1.986	**1.147**	**1.652**
U-Net	80.47	78.35	74.39	76.52	89.15	87.78	84.99	86.63	0.246	0.311	0.318	0.385	1.004	2.085	1.292	1.815
Segformer3D	65.51	66.80	55.42	61.56	79.04	79.84	70.71	76.06	0.656	0.493	1.026	0.741	3.855	2.859	7.739	3.859

TABLE IV
BEAM ABLATION STUDY RESULTS. *Edge*: EDGE ENHANCEMENT, *Attn*: SPATIAL ATTENTION.

Tissue	Edge	Attn	IOU (%) ↑	DSC (%) ↑	ASSD ↓	HD95 ↓
FC	✓	✓	**80.96**	**89.45**	**0.2231**	**0.9262**
	✓	×	80.78	89.34	0.2283	0.9628
TC	✓	✓	78.80	88.06	**0.2788**	**1.8923**
	✓	×	**78.94**	**88.13**	0.2898	2.1175
PC	✓	✓	**75.95**	**85.99**	**0.2618**	**1.0954**
	✓	×	75.36	85.58	0.2683	1.1707
M	✓	✓	**76.96**	**86.88**	0.3682	**1.7438**
	✓	×	76.86	86.83	**0.3670**	1.8180

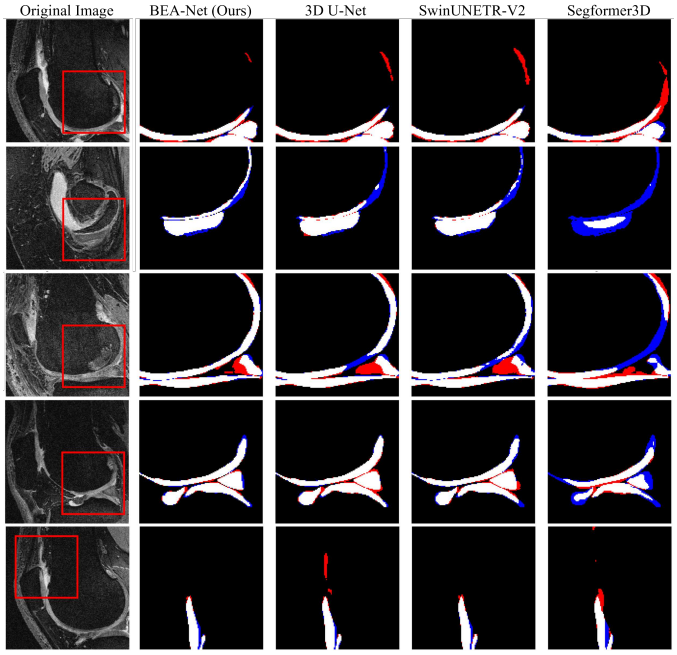

Fig. 2. Example MRI slices with predicted segmentation masks cropped to the red box (left). White: true positives, black: true negatives, blue: false negatives, red: false positives.

U-Net (31.19M), and Segformer3D (4.49M). BEA-Net had more parameters than a standard 3D U-Net, but less than SwinUNETR-V2. Segformer3D achieved the highest FPS (200.15), followed by 3D U-Net (21.72), SwinUNETR-V2 (7.01), and BEA-Net (4.07). While BEA-Net's inference speed was slower than other models, MRI acquisition can take several minutes so it should be satisfactory in a clinical setting.

V. CONCLUSIONS

In this work, BEA-Net was introduced; a multi-task network for segmenting knee cartilage from MRI. Motivated by the challenge of segmenting fine cartilage structure, BEA-Net featured a dual-branch architecture with an auxiliary boundary task. Auxiliary task learning was supported by a boundary enhancement attention module, and a combination loss encouraged consistency between task decoders. BEA-Net was evaluated on MRIs from the OAI and demonstrated its effectiveness compared to other SOTA segmentation models in accurately segmenting fine cartilage structure.

REFERENCES

[1] J. D. Steinmetz et al., "Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050," *The Lancet Rheumatology*, 2023.

[2] J. N. Katz et al., "Diagnosis and Treatment of Hip and Knee Osteoarthritis," *JAMA*, 2021.

[3] G. Matharu et al., "Projections for primary hip and knee replacement surgery up to the year 2060," *Annals of The Royal College of Surgeons of England*, 2022.

[4] Ridhma et al., "Review of automated segmentation approaches for knee images," *IET Image Processing*, 2021.

[5] D. L. Pham et al., "Current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, 2000.

[6] O. Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.

[7] F. Ambellan et al., "Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative.," *Medical image analysis*, 2019.

[8] Y. Zhao et al., "Multi-task deep learning for medical image computing and analysis," *Computers in Biology and Medicine*, Feb. 2023.

[9] R. Sun et al., "Survey of Image Edge Detection," *Frontiers in Signal Processing*, Mar. 2022.

[10] K.-N. Wang et al., "SBCNet," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[11] S. Woo et al., "CBAM: Convolutional Block Attention Module," in *ECCV*, 2018.

[12] A. D. Desai et al., "The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge," *Radiology: Artificial Intelligence*, 2021.

[13] Ö. Çiçek et al., "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *MICCAI*, 2016.

[14] Y. He et al., "SwinUNETR-V2," in *MICCAI*, 2023.

[15] S. Perera et al., "SegFormer3D," in *2024 IEEE/CVF CVPR Workshops*, Jun. 2024.