This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

MBEDNet: A Mamba-CNN Network with Dual Edge-Guided Decoders for Segmentation of TibioFemoral Bone and Cartilage

Armaan Patnaik[1*], Ritanshu Prasad[1*], Pranav Hyagreev[1], Dharshan B.[1], Vivek Maik[1], Manojkumar Lakshmanan[1], and Mohanasankar Sivaprakasam[2†]

[1] Healthcare Technology Innovation Centre (HTIC), IIT Madras, India
[2] Dept. of Electrical Engineering, Indian Institute of Technology Madras, India
mohan@ee.iitm.ac.in

Abstract. Preoperative 3D segmentation of the tibiofemoral joint guides image-based Total Knee Arthroplasty (TKA) and supports the design of patient-specific implants. Automated 3D segmentation in TKA still remains challenging due to thin, low contrast cartilage layers between the tibiofemoral bone structures. In this paper, we propose Mamba Edge-Net (MBEDNet), which integrates the local feature extraction of convolution with long dependency modelling of Mamba to effectively and efficiently segment tibiofemoral anatomies, with sharp contrast between bone and cartilage. MBEDNet is lightweight while achieving the state-of-the-art (SOTA) average Dice similarity coefficient (DSC), against architectures that require 1.3 to 5.5 times more parameters. We have also optimized the network to handle severe osteophytes of Kellgren-Lawrence (KL) grade IV. Evaluating our model on the publicly available OAIZIB-CM knee magnetic resonance imaging (MRI) dataset, we achieved DSC scores of 98.46% and 98.64% on the femur and tibia respectively, and 89.03% and 86.29% on the femoral and tibial cartilage respectively. The code is available at https://github.com/Armaanmerc/MBEDNet

Keywords: Knee MRI · Tibiofemoral Joint Segmentation · Cartilage Segmentation · State Space Model · Total Knee Arthroplasty

1 Introduction

Volumetric MRI can be used to track and monitor Osteoarthritis (OA) progression in knee, typically managed with TKA, in which the diseased tibiofemoral joint is restored by replacing it with a prosthetic implant [21]. Accurate performance of TKA procedures relies heavily on precise segmentation of the femur, tibia, and the articular cartilage that surrounds them. Manually segmenting these complex joint lines across image slices is laborious and time consuming, highlighting the need for reliable automated segmentation methods [26].

Convolution-based encoder-decoder approaches such as U-Net and its derived architectures [8, 15, 18, 28] are among the early benchmarks for medical image

[*]Equal contribution. [†]Corresponding author.

segmentation. Due to the local receptive field of volumetric convolutions, these architectures excel at extracting fine details at the voxel-level, such as articular cartilage boundaries. However, they struggle to model the global dependencies needed to capture broader anatomical relationships and spatial alignment between the femur and tibia. Vision Transformers (ViTs) [9, 10], in contrast, capture global context effectively through self-attention but are less adept at modeling fine local detail; the standard ViT formulation also incurs quadratic complexity $\mathcal{O}(N^2)$ with respect to the number of tokens, which becomes prohibitive for high-resolution volumetric inputs. Hierarchical variants such as the Swin Transformer [13] reduce this computational burden through windowed attention, but at the cost of long-range global context. State Space Models (SSMs) [19, 24, 25] have recently emerged as an alternative that offers a global receptive field at linear complexity, positioning them as a computationally efficient substitute for transformers. However, despite their differing mechanisms, both windowed attention and the sequential, state-averaging formulation of SSMs tend to smooth out fine spatial detail, making it difficult to preserve voxel-level structures such as the bone-cartilage interface and to segment precise boundaries. Several studies across medical imaging have shown that combining global and local context leads to stronger feature representations, whether for MRI reconstruction [16] or segmentation [2], reinforcing the case for hybrid architectural designs. Similar segmentation works have been performed in other imaging domains such as Computed Tomography (CT) [7, 14] and ultrasound imaging [4, 11]. To address these complementary limitations, we propose Mamba Edge-Net (MBEDNet), which combines the global dependency modeling of Mamba with the local feature extraction of CNNs. Our main contributions are as follows: **(1)** We propose MBEDNet, a lightweight hybrid CNN-Mamba architecture that achieves SOTA results while preserving fine boundary precision and overall knee geometry. **(2)** We provide clinical validation on KL grade IV subjects, demonstrating that MBEDNet remains robust and reliable even in advanced OA cases.

2 Methodology

MBEDNet directly operates on the full 3D MRI volume. These input MRIs are single-channel, intensity-normalised volumetric scans at its native anisotropic spacing of $0.365 \times 0.365 \times 0.7$ mm. The network trains on sub-volumes of size $64 \times 192 \times 192$ voxels, cropped along all three axes and biased toward cartilage and bone regions so the thin cartilage classes are adequately represented. A KL-grade aware weighted random sampler is added to oversample severe OA cases, ensuring the model's exposure to osteophytic morphology. At inference, the full volume is reconstructed from overlapping 3D patch predictions via Gaussian-weighted sliding-window aggregation. Figure 1 provides an overview of the complete architecture.

A $7 \times 7 \times 7$ convolutional stem with stride 2 maps the input to 32 channels at half spatial resolution. The encoder proceeds through three stages (32, 64, 128 channels). At each stage, the incoming feature map is processed in parallel by a

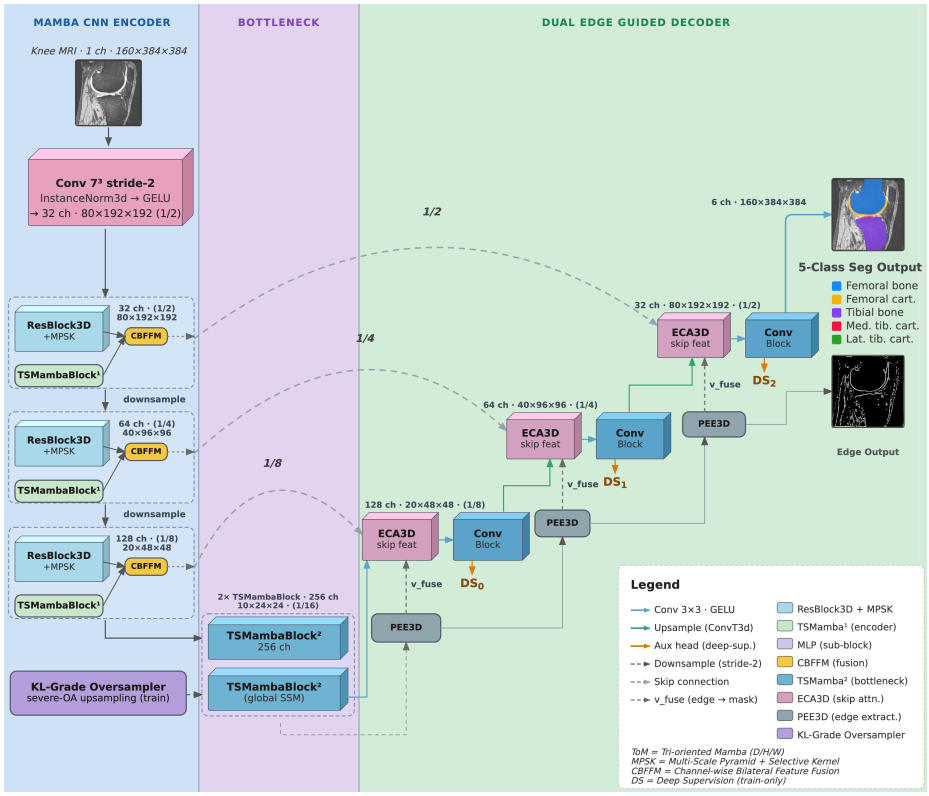


Fig. 1. Overview of the proposed MBEDNet. Encoder runs convolutional (MPSK) and state-space (TSMamba) branches in parallel. The dual decoder jointly predicts segmentation masks and tissue boundaries with vertical cross-task fusion.

Multi-Scale Pyramid Selective Kernel (MPSK) branch [23] and a TSMambaBlock branch [24], and their outputs are merged by a Channel-wise Bilateral Feature Fusion Module (CBFFM) [1]. The fused features serve both as skip connections to the decoders and as input to the next stage, after a stride-2 downsample that doubles the channel count. After the third stage, this yields a 256-channel bottleneck at 1/16 resolution. The bottleneck consists of two stacked TSMambaBlocks that model long-range relations such as relative femur-tibia alignment, which are beyond the reach of a purely local receptive field. Fine boundary detail is reintroduced downstream by the skip connections and edge decoder.

The bottleneck's output initialises two parallel decoders: a semantic mask stream with Efficient Channel Attention (ECA) [22] gated skips, and a boundary stream built around a Progressive Edge Extractor (PEE) [14] that consumes the un-gated encoder skip directly, since ECA's recalibration would attenuate the high-frequency content the extractor depends on. At every decoder level, the boundary stream is injected into the mask stream in one direction only (edge

to mask) via vertical cross-task fusion [14], giving the semantic decoder direct access to interface evidence rather than relying on region overlap alone. The refined mask and edge features are projected to their final output channels and upsampled to native resolution, yielding the five-class segmentation mask and single-channel edge map.

2.1 Mamba-CNN Encoder

A 3D volume of knee MRI data, represented as $X \in \mathbb{R}^{1\times D\times H\times W}$, is first processed by a stem which applies a $7 \times 7 \times 7$ convolution. This increases the volume's channel dimension to 32 and subsamples the volume by a factor of 2 in each dimension, resulting in a feature volume of size $(32, D/2, H/2, W/2)$. The first three encoding stages ($\ell = 1, 2, 3$) contain a TSMamba branch for global context and an MPSK branch for fine local detail, followed by a CBFFM branch.

The TSMamba branch refines the volume three times, with each pass performed along one of the three dimensions (Depth, Height, and Width). Feature propagation along the chosen dimension maintains selective state information, which determines which information is significant to retain or discard. The outputs of the three passes are combined through a $1\times1\times1$ convolution and instance normalized. The resulting tri-oriented Mamba features are given by $F_{\mathrm{TOM}} = \mathrm{IN}\big(\mathrm{Conv}_{1\times1\times1}[S_D(X) \,\|\, S_H(X) \,\|\, S_W(X)]\big)$, where $S_D(X), S_H(X), S_W(X)$ are the selective scans along axis Depth, Height and width respectively, and $\|$ is the channel concatenation operator. Linear complexity is maintained by performing three distinct passes, where the complexity of each pass is $\mathcal{O}(D + H + W)$. This is drastically more efficient than full 3D self-attention, which has quadratic complexity $\mathcal{O}((D \times H \times W)^2)$.

The MPSK branch receives the stage input features F_{in} (stem output for the first stage, and the fused & downsampled features from the previous stage for subsequent stages). This branch applies three parallel 3D convolutions with kernel sizes 1^3, 3^3, and 5^3. A learned per-channel selective kernel mechanism then adaptively weights these multi-scale responses. This enables the network to flexibly use smaller kernels for fine structures (e.g., thin cartilage texture) and larger kernels for broader contextual structures (e.g., bone). Pyramid pooling with grid sizes 2×2, 4×4, and 8×8 provides additional multi-scale contextual information, resulting in F_{pool}. These features are combined using residual connections: $F_{\mathrm{MPSK}} = F_{\mathrm{in}} + F_{\mathrm{SK}} + F_{\mathrm{pool}}$.

The outputs of the two branches are fused using the CBFFM Branch. The features are first concatenated along the channel dimension: $F_{\mathrm{cat}} = [F_{\mathrm{MPSK}} \,\|\, F_{\mathrm{TOM}}]$. This is followed by a $1 \times 1 \times 1$ convolution, GELU activation, and instance normalization, which yields an intermediate representation Z. A Squeeze-and-Excitation (SE) gate then computes channel-wise attention weights: $g = \sigma\big(W_2\,\mathrm{GELU}(W_1\,\mathrm{GAP}(Z))\big)$, where σ is the sigmoid function and GAP denotes global average pooling. The final fused features are computed as $F_{\mathrm{fuse}} = g \odot Z$. The fused features F_{fuse} act as the encoder skip connection S_ℓ for the decoder. They are simultaneously downsampled using a stride-2 convolution that doubles the channel dimension ($32 \to 64 \to 128$).

2.2 Dual Decoder with Vertical Cross-Task Fusion

The bottleneck tensor ($\mathbf{x}_b$) is fed into two level-aligned decoders — a semantic mask stream ($\mathbf{M}$) and a boundary stream ($\mathbf{E}$), decoded over three upscaling stages ($\ell = 1, 2, 3$). Both decoders are initialized from $\mathbf{x}_b$ and interact at every stage through vertical cross-task fusion, injection of boundary-stream features into the mask stream at the same decoder level, so that edge information informs the segmentation output rather than being predicted independently. At every stage ℓ, the semantic mask stream performs a stride-2 transposed convolution ($\mathrm{ConvT}_{2\times}$) on the current feature $\mathbf{M}^{(\ell-1)}$. The matching encoder skip $\mathbf{S}^{(\ell)}$ is recalibrated by an ECA [22] gate, giving the gated skip $\tilde{\mathbf{S}}^{(\ell)}$, which is concatenated with the upsampled feature (denoted $\|$) and refined by two sequential $3 \times 3 \times 3$ blocks, each a convolution followed by $\phi = \mathrm{GELU} \circ \mathrm{IN}$, denoted jointly as $\mathrm{Conv}_3[\cdot]$: $\mathbf{M}^{(\ell)} = \mathrm{Conv}_3\big[\mathrm{ConvT}_{2\times}(\mathbf{M}^{(\ell-1)}) \,\|\, \tilde{\mathbf{S}}^{(\ell)}\big]$.

The boundary stream performs the same $\mathrm{ConvT}_{2\times}$ operation on $\mathbf{E}^{(\ell-1)}$ but concatenates the raw skip $\mathbf{S}^{(\ell)}$, since the edge extractor needs the full high-frequency content that channel gating would attenuate. The result passes through a Progressive Edge Extractor (PEE) [14], which subtracts multi-scale low-pass responses (3D average-pooled, $k \in \{3, 5\}$) from a squeezed feature to expose scale-specific unsharp detail, yielding $\mathbf{E}^{(\ell)} = \mathrm{PEE}\big(\mathrm{ConvT}_{2\times}(\mathbf{E}^{(\ell-1)}) \,\|\, \mathbf{S}^{(\ell)}\big)$ and biasing the boundary stream toward the low-contrast bone-cartilage interfaces.

The boundary evidence is injected into the semantic stream in one direction only (edge $\rightarrow$ mask), producing a fused feature $\tilde{\mathbf{M}}^{(\ell)} = \mathrm{GELU}\big(\mathrm{Conv}_1[\mathbf{M}^{(\ell)} \,\|\, \mathbf{E}^{(\ell)}]\big)$, distinct from $\mathbf{M}^{(\ell)}$; this fused representation, propagates forward as the input to stage $\ell + 1$. The semantic stream stays boundary-aware at every stage — a property that matters most for cartilage spanning two to three voxels, where region overlap alone is ambiguous. Each stage emits a deep-supervision [12] $\mathbf{DS}_\ell = \mathrm{Conv}_1(\tilde{\mathbf{M}}^{(\ell)})$, upsampled to full resolution. After the final stage ($\ell = 3$), $\tilde{\mathbf{M}}^{(3)}$ and $\mathbf{E}^{(3)}$ are projected by $1\times1\times1$ convolutions to C-class and single-channel logits and $\times2$ interpolated to compensate for the stride-2 stem, producing the predicted mask ($\hat{\mathbf{Y}}_{\mathrm{mask}}$) and edge mask ($\hat{\mathbf{Y}}_{\mathrm{edge}}$) at native resolution.

2.3 Cartilage-Aware Composite Loss

MBEDNet was trained using three loss functions. The semantic stream in the dual decoders uses a class-weighted Dice loss ($\mathcal{L}_{Dice}$) [20] together with a distance-weighted cross-entropy loss ($\mathcal{L}_{CE}$) [17], with a scheduling parameter α controlling their tradeoff over training. Dice weights are fixed per class to penalize errors on the smaller cartilage classes more heavily. The distance weights for $\mathcal{L}_{CE}$ are computed dynamically per sample using a Euclidean Distance Transform (EDT), which exponentially upweights voxels located directly at the bone-cartilage interface. The boundary decoder uses edge loss ($\mathcal{L}_{Edge}$ with $\lambda_E = 0.1$) [14] supervised by ground truth edges generated internally by morphological max pooling and erosion. We employed deep supervision ($\mathcal{L}_{deep}$) [12] on the intermediate regional decoder output with decaying weights to prevent gradient dilution across the

upscaling stages. The full objective is defined as:

$$\mathcal{L} = \alpha\mathcal{L}_{Dice} + (1-\alpha)\mathcal{L}_{CE} + \lambda_E\mathcal{L}_{Edge} + \mathcal{L}_{deep} \quad (1)$$

3 Results and Discussion

3.1 Dataset and Implementations:

We trained MBEDNet on the OAIZIB-CM [27] dataset, containing 507 knee MRIs, with expert annotated masks for five classes, each classified by expert radiologists into different KL grades. The 507 volumes are split into 65/15/20 in five folds (101 - 102 test cases each), with every case appearing in exactly one test fold. We ran all experiments along with a comprehensive five fold validation in our model on an A100 80 GB GPU for 200 epochs. AdamW was used as optimizer, setting the learning rate at 1×10^{-4} with cosine annealing down to 1×10^{-6}. Due to memory limits, the per-step batch was kept at 2 with gradient accumulation over 4 steps, giving us an effective batch of 8. Throughout our evaluation, we report DSC for femoral bone (FB), tibial bone (TB), femoral cartilage (FC), and tibial cartilage (TC), where TC is the average of the medial (MTC) and lateral (LTC) tibial cartilage scores.

3.2 Comparative Results:

MBEDNet was evaluated against SOTA baselines trained on OAIZIB and reported in [23] (200 epochs, 5-fold cross-validation). These baseline scores used a 70/20/10 train/validation/test split, while our evaluation uses a 65/15/20 split. MBEDNet and SwinUNETR were each trained on OAIZIB-CM under the same protocol with separate independent splits, under comparable evaluation conditions. As shown in Table 1, MBEDNet matches the larger MSPF-VM-Unet with 23% fewer parameters, and its cartilage-focused training improves cartilage segmentation by roughly 1-2% over SwinUNETR. This indicates that boundary-aware Mamba encoders capture thin-structure geometry more effectively than pure transformer or heavier Mamba-CNN architectures. We did not perform significance testing here, as the baseline scores are reproduced from [23] rather than recomputed on our test set; without per-case predictions under a matched protocol, the paired, fold-dependence-aware inference required for a valid comparison is not possible.

3.3 Ablation Analysis:

Table 2 quantifies the contribution of each MBEDNet component. Removing the Mamba branch causes the largest drop, confirming that global context is essential for joint segmentation, while removing the edge decoder degrades HD95, reflecting its role in preserving thin cartilage boundaries. Performance therefore arises from the synergy of all components rather than any single one. Per-KL-grade results (Table 3) show consistently high scores across grades, with a modest decline at

Table 1. Comparison of segmentation performance across the entire test set (DSC, %). Baseline results (mean from 5-fold CV) are reported from [23]; MBEDNet is evaluated under 5-fold cross-validation.

Method	Params	FB	TB	FC	TC	Avg
U-Net [18]	31.0 M	96.43	96.05	84.73	83.21	90.11
V-Net [15]	71.0 M	97.04	96.52	87.75	84.09	91.35
nnU-Net [8]	30.0 M	97.87	97.17	88.51	85.49	92.26
I-Mask-RCNN [5]	44.0 M	97.60	97.53	83.32	80.67	89.78
SwinUNETR [6]	62.2 M	98.46	98.47	87.77	84.47	92.29
MtRA-Unet [3]	96.5 M	98.53	98.44	89.19	86.02	93.05
VM-UNet [19]	27.0 M	97.35	96.31	85.33	83.17	90.54
MSPF-VM-Unet [23]	22.7 M	**98.61**	98.37	**89.57**	85.63	93.05
MBEDNet (ours)	**17.4** M	98.46	**98.64**	89.03	**86.29**	**93.11**

Table 2. Ablation study of MBEDNet components on five classes. HD95 – 95th percentile Hausdorff distance, ASSD – Average Symmetric Surface Distance. All the numbers reported are averaged over five classes.

Configuration	DSC ↑ (%)	HD95 ↓ (mm)	ASSD ↓ (mm)
w/o Mamba branch	89.80	1.44	0.400
w/o Edge decoder	90.70	1.29	0.360
w/o Bottleneck (Mamba→conv)	91.00	1.14	0.330
w/o Edge loss	91.60	1.03	0.310
Full model	92.00	0.95	0.289

KL Grade IV, where cartilage degradation is most pronounced and small spurious voxels occasionally appear near severely eroded interfaces. This indicates that MBEDNet generalizes reliably across disease severity. Model differences were assessed on per-case Dice scores using a two-sided Wilcoxon signed-rank test, paired across the test fold cases. For the per-structure comparisons, p-values were corrected for multiple comparisons using the Holm–Bonferroni procedure; corrected $p < 0.05$ was considered significant.

3.4 Clinical Validation:

We assessed MBEDNet predicted masks surgically to validate surface geometry, anatomical landmark consistency, surface registration and surgical tracking on KL grade IV femoral segmentation. Predicted and ground-truth masks were converted into triangular meshes (refer to Figure 2), an inhouse anatomical landmark annotation module was used to localize eight clinically relevant femoral landmarks, across five KL Grade IV test cases. The average of mean Euclidean error calculated between predicted and ground-truth mesh landmarks was 0.964 ± 0.528 mm. A randomised test subject was 3D-printed as physical phantom and

Table 3. Segmentation Performance Across Files with KL Grades (DSC, %). Out of 102 test files, 7 did not report a KL grade.

KL	n	FB	TB	FC	TC	avg DSC	HD95	ASSD
0	20	98.76	98.82	88.62	85.99	93.05	0.73	0.149
1	12	98.74	98.71	89.74	88.06	93.81	0.70	0.141
2	21	98.59	98.62	88.48	86.92	93.15	0.79	0.151
3	27	98.58	98.66	88.88	84.84	92.74	0.82	0.167
4	15	98.64	98.39	88.77	82.93	92.18	1.07	0.238
All	95	98.66	98.64	88.90	85.75	92.99	0.95	0.289

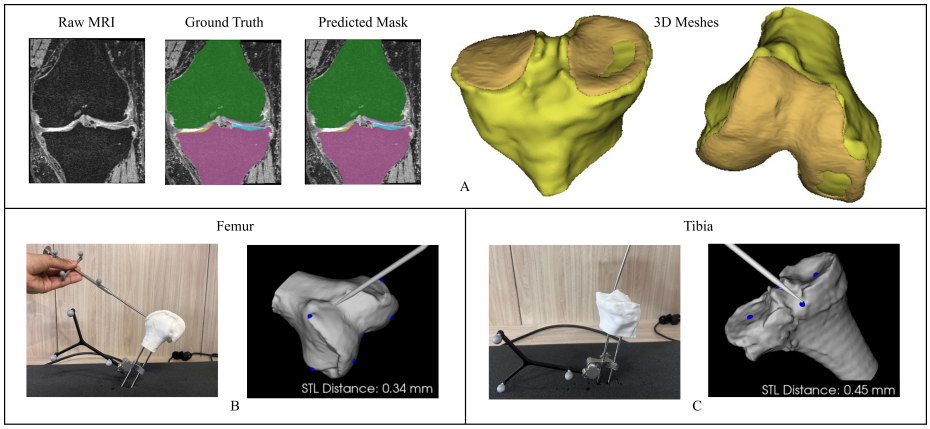


Fig. 2. (A) Sample result of the segmentation masks and their corresponding 3D meshes and (B,C) Phantom tracking of Femur and Tibia respectively.

evaluated using real-time optical tracking. Predicted landmarks were registered to manually annotated landmarks in the DRB space using an IR stereo camera and intraoperative blunt probe. The target registration error (TRE) between the predicted and annotated landmarks, collected independently of the registration, was 0.045 mm. The segmented STL model was transformed using this registration, and the tracked tool-tip distance to the registered bone surface was 0.395 mm, falling well within reported navigation tolerances.

4 Conclusion and Future Works

This paper introduces MBEDNet, a highly efficient Mamba-CNN architecture developed to accurately segment tibiofemoral joints for image-based TKA workflows. MBEDNet performs competitively with existing SOTA architectures with a lower computational cost, providing high memory efficiency and rapid processing. Our model remains robust even in advanced disease stages (KL grade IV), assisting surgeons with accurate preoperative segmentation masks regardless of disease

severity. Future works could be focused on further noise reduction, improving the model for cartilage tissue segmentation. The architecture can be adapted for CT scans to support more comprehensive surgical imaging planning.

Acknowledgments. This work was supported by the Ministry of Electronics and Information Technology (MeitY), Government of India.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Atabansi, C.C., Wang, S., Li, H., Nie, J., Xiang, L., Zhang, C., Liu, H., Zhou, X., Li, D.: DCM-Net: dual-encoder CNN-Mamba network with cross-branch fusion for robust medical image segmentation. BMC Medical Imaging **25**(1), 395 (2025)
2. 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)
3. Daydar, A., Pramanick, A., Sur, A., Kanagaraj, S.: Segmentation of tibiofemoral joint tissues from knee MRI using MtRA-Unet and incorporating shape information: Data from the osteoarthritis initiative. arXiv preprint arXiv:2401.12932 (2024)
4. Du Toit, C., Orlando, N., Papernick, S., Dima, R., Gyacskov, I., Fenster, A.: Automatic femoral articular cartilage segmentation using deep learning in three-dimensional ultrasound images of the knee. Osteoarthritis and Cartilage Open **4**(3), 100290 (2022)
5. Felfeliyan, B., Hareendranathan, A., Kuntze, G., Jaremko, J.L., Ronsky, J.L.: Improved-Mask R-CNN: Towards an accurate generic MSK MRI instance segmentation platform (data from the osteoarthritis initiative). Computerized Medical Imaging and Graphics **97**, 102056 (2022)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv preprint arXiv:2201.01266 (2022)
7. Humayun, A., Rehman, M., Liu, B.: A method framework of semi-automatic knee bone segmentation and reconstruction from computed tomography (CT) images. Quantitative Imaging in Medicine and Surgery **14**(10), 7151–7175 (2024)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
9. Kakavand, R., Palizi, M., Tahghighi, P., Ahmadi, R., Gianchandani, N., Adeeb, S., Souza, R., Edwards, W.B., Komeili, A.: Integration of Swin UNETR and statistical shape modeling for a semi-automated segmentation of the knee and biomechanical modeling of articular cartilage. Scientific Reports **14**, 2748 (2024)
10. Kakavand, R., Tahghighi, P., Ahmadi, R., Edwards, W.B., Komeili, A.: Swin UNETR segmentation with automated geometry filtering for biomechanical modeling of knee joint cartilage. Annals of Biomedical Engineering **53**(4), 908–922 (2025)
11. Kompella, G., Antico, M., Sasazawa, F., Jeevakala, S., Ram, K., Fontanarosa, D., Pandey, A.K., Sivaprakasam, M.: Segmentation of femoral cartilage from knee ultrasound images using mask R-CNN. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 966–969. IEEE (2019)

12. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022. IEEE (2021)
14. Marsilio, L., Faglia, A., Rossi, M., Mainardi, L., Manzotti, A., Cerveri, P.: CEL-Unet: a novel CNN architecture for 3D segmentation of knee bones affected by severe osteoarthritis for PSI-based surgical planning. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE (2022)
15. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016). https://doi.org/10.1109/3DV.2016.79
16. Murugesan, B., Vijaya Raghavan, S., Sarveswaran, K., Ram, K., Sivaprakasam, M.: Recon-GLGAN: A global-local context based generative adversarial network for MRI reconstruction. In: Machine Learning for Medical Image Reconstruction (MLMIR), held in conjunction with MICCAI 2019. Lecture Notes in Computer Science, vol. 11905, pp. 3–15. Springer (2019)
17. Polat, G., Ergenc, I., Kani, H.T., Alahdab, Y.O., Atug, O., Temizel, A.: Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In: Annual Conference on Medical Image Understanding and Analysis. pp. 157–171. Springer (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Ruan, J., Li, J., Xiang, S.: VM-UNet: Vision mamba UNet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
20. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. LNCS, vol. 10553, pp. 240–248. Springer (2017)
21. Varacallo, M.A., Luo, T.D., Mabrouk, A., Johanson, N.A.: Total knee arthroplasty techniques. In: StatPearls [Internet]. StatPearls Publishing, Treasure Island (FL) (May 2024), updated 2024 May 6
22. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
23. Wang, X., Fu, Y., Lu, H., Xia, Y., Cai, X.: VM-Unet enhanced with multi-scale pyramid feature extraction for segmentation of tibiofemoral joint tissues from knee MRI. PLoS One **20**(8), e0330740 (2025)
24. Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-UNet: UNet-like pure visual mamba for medical image segmentation. arXiv preprint arXiv:2402.05079 (2024)
25. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 578–588. Springer, Cham (2024)

26. Yang, M., Colak, C., Chundru, K.K., Gaj, S., Nanavati, A., Jones, M.H., Winalski, C.S., Subhas, N., Li, X.: Automated knee cartilage segmentation for heterogeneous clinical MRI using generative adversarial networks with transfer learning. Quantitative Imaging in Medicine and Surgery **12**(5), 2620–2633 (2022). https://doi.org/10.21037/qims-21-459
27. Yao, Y., Zhong, J., Zhang, L., Khan, S., Chen, W.: CartiMorph: A framework for automated knee articular cartilage morphometrics. Medical Image Analysis **91**, 103035 (2024). https://doi.org/10.1016/j.media.2023.103035
28. 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. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1