Traumatology and Orthopaedics of Kazakhstan
ISSN - e2789-9640

Original article

Advancing Orthopaedic Imaging: Self-Supervised Vision Transformers for 3D Reconstruction of Knee Magnetic Resonance Imaging

Amanali Bekbolat [1], Kassymbek Ozhikenov [2], Roza Beisembekova [3], Damira Mussina [4*], Chingiz Alimbayev [5], Nariman Imashev [6]

[1] Research Assistant of the Robotics Department, Nazarbayev University, Astana, Kazakhstan
[2] Head of the Department of Robotics and technical means of automation, Kazakh National Research Technical University named after K.I. Satbayev, Almaty, Kazakhstan
[3] Associate Professor of the Department of Software Engineering, Kazakh National Research Technical University named after K.I. Satbayev, Almaty, Kazakhstan
[4] Research Assistant, Nazarbayev University, Astana, Kazakhstan
[5] Associate Professor, Department of Robotics and technical means of automation, Institute of Automation and Information Technologies, Kazakh National Research Technical University named after K.I. Satpayev, Almaty, Kazakhstan
[6] Research Assistant of the Robotics Department, Nazarbayev University, Astana, Kazakhstan
***Corresponding author**: damira.pernebayeva@nu.edu.kz

Received: 14 October 2025
Revised: 24 November 2025
Accepted: 12 December 2025
Published: 30 December 2025

Citation: Amanali Bekbolat, Kassymbek Ozhikenov, Roza Beisembekova, Damira Mussina, Chingiz Alimbayev, Nariman Imashev. Advancing Orthopaedic Imaging: Self-Supervised Vision Transformers for 3D Reconstruction of Knee Magnetic Resonance Imaging. Trauma & Ortho Kaz, 2025, 76 (6), jto027. https://doi.org/10.52889/1684-9280-2025-76-6-jto027

Abstract

The main aim of this research is to develop and evaluate a self -supervised learning model based on Vision Transformers for three-dimensional reconstruction of knee Magnetic Resonance Imaging. Three-dimensional reconstruction of hard tissues is an essential procedure for visualization and structural analysis of bones, both in clinical and research settings. Traditional compressed sensing approaches rely on handcrafted priors and iterative solvers, while supervised convolutional neural networks require large datasets of fully sampled Magnetic Resonance Images, which are costly to obtain. To address these limitations, this paper proposes the self-supervised framework based on Vision Transformers architecture along with the integrated convolutional neural network decoder for reconstructing volumetric information from an under-sampled Magnetic Resonance Imaging data. We propose a self-supervised learning framework using Vision Transformers for three – dimensional knee magnetic resonance image reconstruction. Our method leverages masked patch pre-training on a custom dataset of proton density fat-saturated knee magnetic resonance imaging scans, followed by volumetric decoding with a lightweight three- dimensional convolutional neural network. Results demonstrate competitive reconstruction quality with Dice score of 0.87, IoU of 0.83, and recall of 0.93, showing that self-supervised vision transformers effectively capture long-range dependencies and volumetric continuity. This study highlights the potential of Vision transformer-based self – supervised learning frameworks to overcome the limitations of traditional supervised deep learning approaches. The proposed method demonstrated that it can reduce dependence on large fully sampled datasets and support accurate three -dimensional reconstructions for clinical and research applications.

Keywords: knee joint, Magnetic Resonance Imaging, 3D Imaging, anterior cruciate ligament, image segmentation, transformers, Artificial Intelligence.

1. Introduction

Magnetic Resonance Imaging (MRI) of the knee provides high-resolution, contrast-rich 3D views of joint tissues, but long scan times pose a major challenge. While MRI inherently acquires volumetric information, clinical images are typically examined slice by slice, and accurate 3D reconstruction is crucial to transform these slices into volumetric representations that enable precise diagnosis, surgical planning, and treatment monitoring. Beyond clinical use, reconstructed volumes are indispensable in biomechanics simulations, robotic-assisted surgery, and personalized implant design, providing quantitative and anatomically faithful models that 2D imaging alone cannot deliver [1].

Achieving such reconstructions efficiently is challenging. Accelerating MRI by under-sampling k-space leads to an ill-posed inverse problem rife with aliasing artifacts [2]. Traditional compressed sensing (CS) methods introduced sparsity priors to recover images but require lengthy iterative solvers and careful tuning [3]. Since the late 2010s, deep learning has revolutionized MRI reconstruction by learning inverse mappings directly from data, achieving faster and higher-quality results than CS [2,4-10]. Initiatives such as the NYU fast MRI dataset and SKM-TEA have spurred numerous approaches. Unlike prior studies that rely solely on public datasets, this work leverages a custom high-resolution knee MRI dataset, enabling evaluation under conditions closer to real-world practice where heterogeneity and limited fully sampled scans remain challenges [2].

In this paper, we present a self-supervised learning framework based on ViTs for three-dimensional knee MRI reconstruction, designed to capture long-range anatomical dependencies, mitigate reliance on fully sampled datasets, and move toward clinically applicable accelerated imaging solutions.

2. Materials and methods

The dataset utilized in this study comprises Proton Density (PD) fat-saturated MRI scans of the human knee, acquired in axial, coronal, and sagittal planes, in NRRD format. Ground truth volumes for training were obtained through manual 3D reconstruction using 3D Slicer software. The reconstruction process involved the Grow from Seeds segmentation method, which allows region-based expansion from annotated seed points. To refine the anatomical boundaries and ensure structural completeness, Gaussian smoothing was applied to reduce noise, followed by morphological operations such as hole closing to fill discontinuities in segmented structures. The examples of this segmentation are illustrated in Figure 1 (A-D), which show axial, coronal, and sagittal MRI slices overlaid with the segmentation mask, as well as a 3D rendering of the reconstructed knee anatomy.

The pre-processing steps include normalization to [0,1] range, resampling to consistent voxel spacing, followed by patch-wise cropping and padding. The augmentation techniques, such as random flipping, affine warping, Gaussian noise, and cutmix masking, were applied. Preprocessing included normalization:

$$I' = \frac{I - \mu}{\sigma},$$

resampling to isotropic voxel spacing, and patch-wise cropping/padding. Data augmentation included random flipping, affine warping, Gaussian noise, and cutmix masking.

After acquiring the annotated dataset, all volumes were pre-processed before the training procedure. Scans were resampled to a fixed resolution of 320 x 320 x 36 using bilinear interpolation to ensure consistency across the dataset.

https://doi.org/10.52889/1684-9280-2025-76-6-jto027

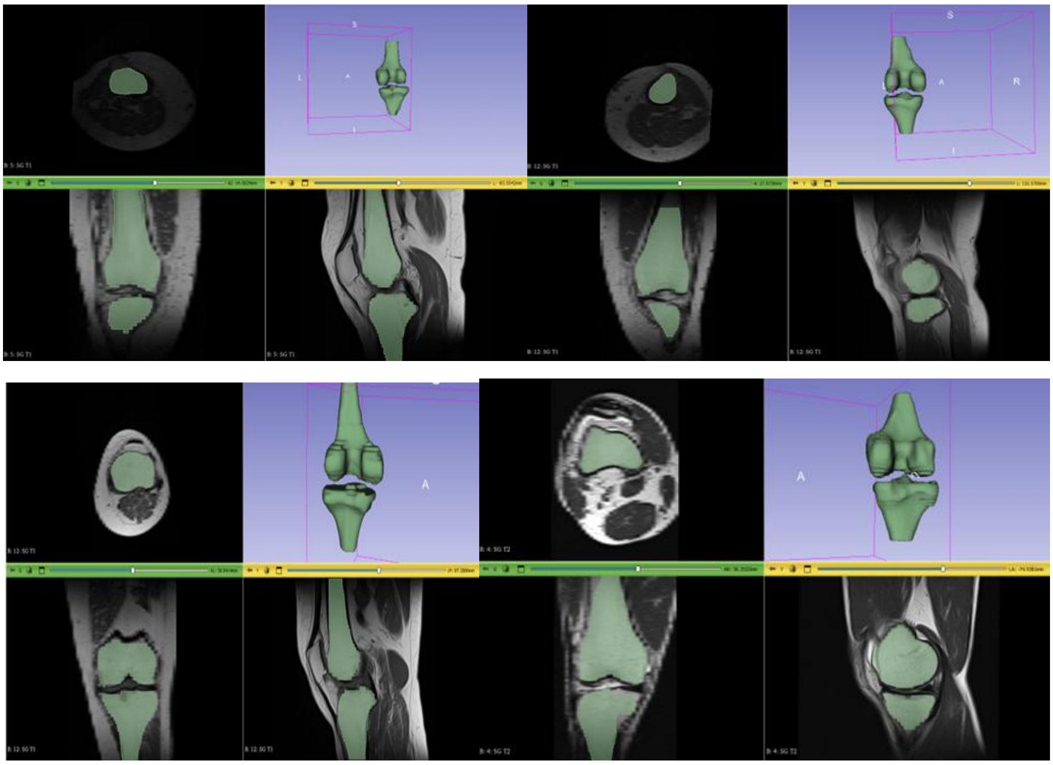


Figure 1 – Manual 3D -Reconstruction of knee joint hard tissues in 3D Slicer software

The architecture is a Vision Transformer model that processes volumetric MRI data as sequences of embedded 3D patches. Each volume is divided into non-overlapping patches using a 3D convolutional embedding layer, and the resulting patch tokens are passed through a stack of Transformer encoder blocks. The encoded representations are then projected back to voxel intensities through a linear decoder, and reconstructed patches are reassembled into a coherent 3D volume. Model training is performed end-to-end using mean squared error between the input and reconstructed volumes, with an optional SSIM–MSE combined loss.

Patch embedding was implemented through the 3D convolutional layer with kernel and stride equal to the patch dimensions. Input slices are divided into patches, projected as:

$$x_p = Flatten(P_i)W_e + E_{pos},$$

where W_e is the learnable embedding matrix and positional encodings. In its core, ViT relies on self-attention to capture long-range dependencies across MRI patches. Given a set of embedded tokens $\{x_1, x_2.....x_n\}$ queries (Q), keys (K), and values (V) are generated through learned linear projections. The self-attention mechanism is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-head attention extends this mechanism to multiple representation subspaces, improving the ability of the network to learn both global and local features of knee MRI volumes.

The dataset was split into training and validation subsets with an approximate 80:20 ratio. The model was trained to minimize reconstruction error using a mean squared error (MSE) loss between the predicted 3D volume and the ground-truth MRI volume.

Optimization was performed using Adam with an initial learning rate of 1e-4 and weight decay regularization to prevent overfitting. Training was conducted over 100 epochs with a batch size tuned to available GPU memory.

The training and validation loss curves (Figure 2) show rapid convergence within the first 20 epochs, followed by gradual improvements in later epochs. Both curves plateaued at low values (~0.1), indicating stable generalization without significant overfitting.

Evaluation involved both qualitative and quantitative analyses:

- Dice coefficient.
- Intersection over Union (IoU).
- Hausdorff Distance (HD95).
- Visual overlays comparing predicted masks with expert annotations (Figure 1).
- Training convergence curves.

3. Results

Quantitative Results

The validation performance of the proposed ViT-based reconstruction model is summarized in Table 1. The model demonstrates strong reconstruction accuracy across all evaluated metrics, indicating its robustness and reliability on the validation dataset.

Table 1 – Validation Performance Metrics of the ViT-Based Reconstruction Model

Metric	Value
Dice Score	0.8727
IoU	0.8385
F1 Score	0.8727
Recall	0.9290

Training Curves

The evolution of the training and validation loss over 100 epochs is depicted in Figure 2. The consistent decline in both curves, along with their convergence around epoch 50, indicates stable learning and good generalization. The training and validation loss curves show rapid convergence within the first 20 epochs, followed by gradual improvements in later epochs. Both curves plateaued at low values (~0.1), indicating stable generalization without significant overfitting. The absence of significant overfitting confirms the effectiveness of the model's self-supervised learning setup and the adequacy of the applied regularization and data augmentation strategies.

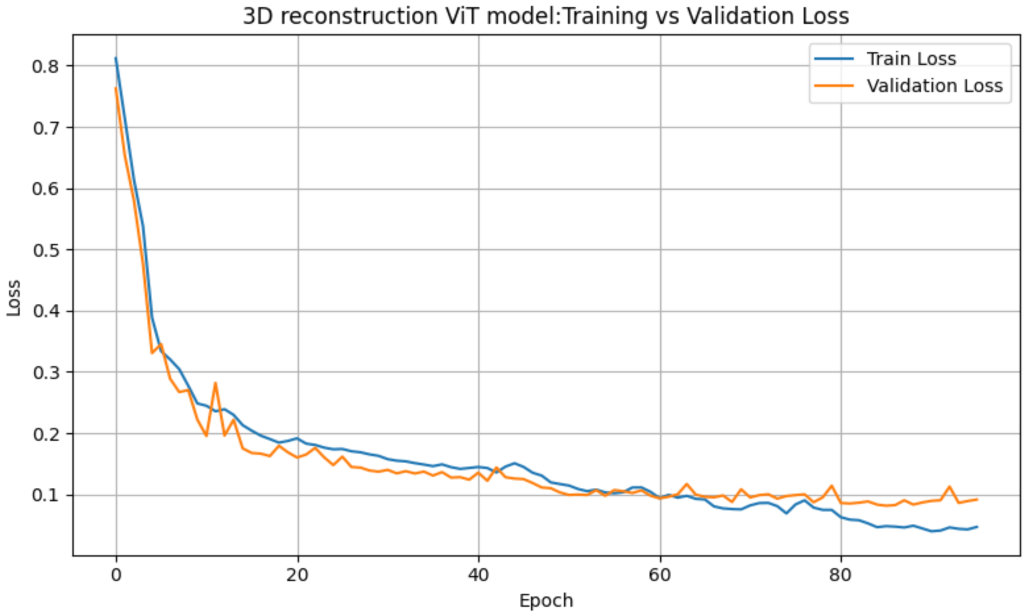
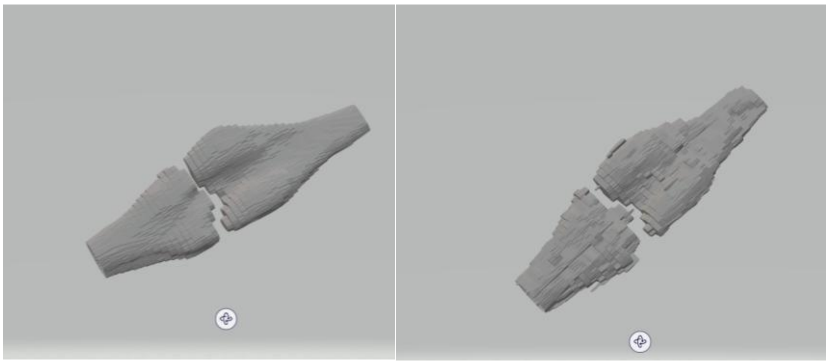


Figure 2 – Training and Validation Loss Curve of the ViT Model for 3D Reconstruction

Qualitative Results

Following training, the model outputs were reconstructed into 3D voxel grids. The voxelized outputs were rendered into surface meshes to visualize anatomical structures such as femoral and tibial bone segments (Figure 3). These reconstructions demonstrate that the ViT model successfully captured volumetric continuity from 2D MRI slices, yielding plausible anatomical 3D representations.

https://doi.org/10.52889/1684-9280-2025-76-6-jto027



Figures 3 – Comparison of the original (A) obtained from MRI images and the predicted (B) STL models of the knee joint obtained via a self-supervies VIT

Vision Transformer Experiments

Preliminary ViT results demonstrated capacity for volumetric feature extraction.

The findings are summarized in the Table 2.

Table 2 – Preliminary Vision Transformer performance

Model	Reconstruction Quality	Observations
Baseline ViT	Good spatial consistency	Captures long-range dependencies
Hybrid ViT-CNN	Improved structure recovery	Promising for ACL-focused tasks

4. Discussion

Our experiments demonstrate that ViTs provide a strong foundation for 3D knee MRI reconstruction, capturing long-range contextual relationships across slices that conventional CNNs often fail to model. By integrating a self-supervised masked patch pretraining strategy with a lightweight 3D CNN decoder, the framework effectively balances global attention with local structural recovery. This combination aligns with recent findings that hybrid CNN Transformer pipelines can outperform CNNs alone, offering improved volumetric continuity without requiring large fully sampled training datasets.

At the same time, several limitations must be acknowledged. First, the dataset used in this study, while carefully curated, remains modest in size and restricted to proton density fat-saturated sequences. The absence of external validation across diverse scanners and institutions limits generalizability. Second, although quantitative results were competitive, evaluation relied primarily on Dice, IoU, and Recall metrics. Standard image quality measures such as PSNR and SSIM, as well as ablation studies on masking strategies and model depth, were not explored in this work but could provide deeper insights. Finally, the computational demands of ViTs remain a challenge for deployment in clinical environments, highlighting the need for lightweight architectures and benchmarking of inference efficiency.

Despite these constraints, the findings underscore the clinical promise of SSL-based ViTs. By enabling accurate volumetric reconstructions from undersampled acquisitions, the method has the potential to accelerate MRI workflows, reduce patient burden, and support advanced applications such as biomechanics simulations and surgical planning. Addressing the identified limitations through larger multi-institutional studies, richer benchmarking, and optimized hybrid architectures will be important steps toward clinical translation.

5. Conclusions

This study introduced a self-supervised Vision Transformer framework for 3D knee MRI reconstruction, combining masked patch pretraining with a lightweight 3D convolutional decoder. The proposed approach demonstrated strong performance, achieving a Dice score of 0.87, IoU of 0.83, and Recall of

https://doi.org/10.52889/1684-9280-2025-76-6-jto027

0.93, while effectively capturing both local and global dependencies across MRI volumes. These findings highlight the ability of SSL-ViTs to reduce reliance on fully sampled datasets, offering a practical pathway toward accelerated and clinically reliable reconstructions.

Beyond methodological contributions, the results underscore the clinical significance of rapid, artifact-free volumetric reconstructions for diagnosis, surgical planning, and biomechanics applications. By bridging efficiency and accuracy, this approach can contribute to improving diagnostic confidence, reducing patient scan times, and supporting advanced orthopedic research.

Future directions will focus on expanding evaluation with additional reconstruction benchmarks, refining hybrid ViT–CNN architectures to preserve fine details while reducing computational demands, and validating models on multi-institutional datasets. These steps will be critical for translating the proposed framework into real-world clinical practice.

Conflict of interests. The authors declare no conflict of interest.

Acknowledgements. The authors thank the Ministry of Science and Higher Education of the Republic of Kazakhstan for supporting this study.

Funding. This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24992820

Author contributions: Conceptualization: A.B., K.O., Methodology: A.B., N.I., Ch.A., Data curation: A.B.; Software: A.B., N.I., Ch.A., Validation: A.B., Writing—original draft: D.M., A.B.; Writing—review & editing: D.M., R.B.; Supervision: K.O., R. B.

References

1. Liu, T., Lu, Y., Xu, J., Yang, H., & Hu, J. (2024). 3D reconstruction of bone CT scan images based on deformable convex hull. Medical & biological engineering & computing, 62(2), 551–561. https://doi.org/10.1007/s11517-023-02951-7

2. Meng, Y., Yang, Z., Shi, Y., & Song, Z. (2025). Boosting vit-based mri reconstruction from the perspectives of frequency modulation, spatial purification, and scale diversification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 6, pp. 6135-6143). https://doi.org/10.1609/aaai.v39i6.32656

3. Wang, A., McDonagh, S., & Davies, M. (2025). Benchmarking Self-Supervised Learning Methods for Accelerated MRI Reconstruction. arXiv preprint arXiv:2502.14009. https://doi.org/10.48550/arXiv.2502.14009

4. Yaman, B., Hosseini, S. A. H., Moeller, S., Ellermann, J., Uğurbil, K., & Akçakaya, M. (2020). Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. Magnetic resonance in medicine, 84(6), 3172–3191. https://doi.org/10.1002/mrm.28378

5. Wang, S., Su, Z., Ying, L., Peng, X., Zhu, S., Liang, F., Feng, D., & Liang, D. (2016). Accelerating magnetic resonance imaging via deep learning. Proceedings. IEEE International Symposium on Biomedical Imaging, 2016, 514–517. https://doi.org/10.1109/ISBI.2016.7493320

6. Huang, P., Li, H., Liu, R., Zhang, X., Li, X., Liang, D., & Ying, L. (2023). Accelerating MRI Using Vision Transformer with Unpaired Unsupervised Training. Proceedings of the International Society for Magnetic Resonance in Medicine ... Scientific Meeting and Exhibition. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition, 31, 2933.

7. Sriram, A., Zbontar, J., Murrell, T., Defazio, A., Zitnick, C. L., Yakubova, N., ... & Johnson, P. (2020). End-to-end variational networks for accelerated MRI reconstruction. In International conference on medical image computing and computer-assisted intervention (pp. 64-73). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-59713-9_7

8. Zhang, C., Chen, S., Cigdem, O., Rajamohan, H. R., Cho, K., Kijowski, R., & Deniz, C. M. (2025). MR-Transformer: A Vision Transformer-based Deep Learning Model for Total Knee Replacement Prediction Using MRI. Radiology. Artificial intelligence, 7(5), e240373. https://doi.org/10.1148/ryai.240373

9. Lin, K., & Heckel, R. (2022). Vision transformers enable fast and robust accelerated MRI. In International Conference on medical imaging with deep learning (pp. 774-795). PMLR.

10. Guo, P., Mei, Y., Zhou, J., Jiang, S., & Patel, V. M. (2023). Reconformer: Accelerated mri reconstruction using recurrent transformer. IEEE transactions on medical imaging, 43(1), 582-593. https://doi.org/10.1109/TMI.2023.3314747

Ортопедиялық бейнелеуді жетілдіру: Тізе буынының магнитті-резонанстық томографиясын 3D реконструкциялауға арналған өзін-өзі қадағалайтын көру трансформерлері

Бекболат А. [1], Ожикенов К.А. [2], Бейсембекова Р.Н. [3],
Мусина Д. [4], Алимбаев Ч.А. [5], Имашев Н. [6]

[1] Робототехника департаментінің ғылыми қызметкері, Назарбаев Университеті, Астана, Қазақстан
[2] Робототехника және автоматтандырудың техникалық құралдары кафедрасының меңгерушісі, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
[3] Бағдарламалық инженерия кафедрасының доценті, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
[4] Ғылыми қызметкер, Назарбаев Университеті, Астана, Қазақстан
[5] Қауымдастырылған профессор, Робототехника және автоматика техникалық құралдары кафедрасы, Автоматика және ақпараттық технологиялар институты, Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
[6] Компьютерлік ғылымдар департаментінің ғылыми қызметкері, Назарбаев Университеті, Астана, Қазақстан

Түйіндеме

Бұл зерттеудің негізгі мақсаты – тізе буынының магнитті-резонанстық бейнелеуін 3D реконструкциялауға арналған Vision Transformers негізіндегі өздігінен білім алу моделін әзірлеу және бағалау. Қатты тіндерді үш өлшемді реконструкциясы клиникалық және ғылыми-зерттеу мақсаттары үшін сүйектерді бейнелеу және құрылымдық талдаудың ажырамас бөлігі болып табылады. Compressed Sensing сияқты дәстүрлі тәсілдер қолмен қалыптастырылатын априорлық деректерге және итерациялық шешушілерге (Iterative solvers) негізделеді, ал бақыланатын конволюционды нейрондық желілер толық таңдалған МРТ деректерінің үлкен көлемін талап етеді, мұндай деректерді жинау айтарлықтай шығынды қажет етеді. Осы шектеулерді шешу үшін бұл мақалада Vision Transformers архитектурасына негізделген өздігінен оқытылатын оқыту үлгісін және жеткіліксіз таңдамалы МРТ деректерінен көлемді қайта құру үшін енгізілген CNN декодерін ұсынады. Біз тізе буынының магниттік-резонанстық кескіндерін 3D реконструкциялау үшін Vision Transformers көмегімен өздігінен жаттығу үлгісін ұсынамыз. Ұсынылған әдіс протондық тығыздықпен май тінін басу режиміндегі тізе буынының МРТ деректер жиынтығына негізделген бетперделенген алдын ала оқытуды пайдаланады, одан кейін жеңілдетілген 3D конволюциялық нейрондық желі көмегімен көлемдік декодтау жүзеге асырылады. Нәтижелер реконструкция сапасының жоғары екенін көрсетеді: Dice коэффициенті – 0,87, IoU – 0,83 және толықтық – 0,93. Бұл өзін-өзі бақылайтын Vision Transformers модельдерінің ұзақ мерзімді тәуелділіктерді және көлемдік үздіксіздікті тиімді түрде меңгере алатынын дәлелдейді. Бұл зерттеу дәстүрлі қадағаланатын терең оқыту әдістерінің шектеулерін еңсеру үшін Vision Transformers негізіндегі өзін-өзі басқаратын оқыту жүйелерінің әлеуетін көрсетеді. Ұсынылған әдіс үлкен, толық таңдалған деректер жиынына тәуелділікті азайта алатындығын және клиникалық және зерттеу қолданбалары үшін дәл 3D реконструкцияларды қамтамасыз ете алатынын көрсетті.

Түйін сөздер: тізе буыны, магниттік-резонансты томография, 3D визуализациясы, алдыңғы айқас байлам, кескіндерді сегментациялау, трансформерлер, жасанды интеллект.

Совершенствование ортопедической визуализации: Самообучающиеся трансформеры зрения для 3D-реконструкции магнитно-резонансная томографии коленного сустава

Бекболат А. [1], Ожикенов К.А. [2], Бейсембекова Р.Н. [3],
Мусина Д. [4], Алимбаев Ч.А. [5], Имашев Н. [6]

[1] Научный сотрудник департамента робототехники, Назарбаев Университет, Астана, Казахстан
[2] Заведующий кафедрой робототехники и технических средств автоматизации, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан

https://doi.org/10.52889/1684-9280-2025-76-6-jto027

[3] Доцент кафедры программной инженерии, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан
[4] Научный сотрудник, Назарбаев Университет, Астана, Казахстан
[5] Ассоциированный профессор, кафедра робототехники и технических средств автоматики, Институт автоматики и информационных технологий, Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан
[2] Научный сотрудник департамента компьютерных наук, Назарбаев Университет, Астана, Казахстан

Резюме

Основной целью данного исследования является разработка и оценка модели самообучения на основе Vision Transformers для трехмерной реконструкции магнитно-резонансной томографии коленного сустава. Трехмерная реконструкция твердых тканей является неотъемлемой процедурой визуализации и структурного анализа костей как в клинических, так и в исследовательских целях. Традиционные подходы как Compressed Sensing основаны на ручном создании априорных данных и Iterative solvers (итерационных решателях), в то время как контролируемые сверточные нейронные сети требуют больших наборов данных с полной выборкой магнитно-резонансных изображений, получение которых является затратным. Для устранения этих ограничений в данной статье предлагается самообучаемая модель обучения на основе архитектуры Vision Transformers вместе со встроенным декодером CNN для реконструкции объемной информации из данных магнитно-резонансной томографии с недостаточной выборкой. Мы предлагаем самообучаемую модель обучения с использованием Vision Transformers для трехмерной реконструкции магнитно-резонансных изображений коленного сустава. Наш метод использует маскированное предварительное обучение на основе набора данных магнитно-резонансной томографии коленного сустава с насыщением жировой ткани протонной плотностью, после чего выполняется объемное декодирование с помощью облегченной трехмерной сверточной нейронной сети. Результаты демонстрируют конкурентоспособное качество реконструкции с оценкой Дайса 0,87, IoU 0,83 и полнотой 0,93, что свидетельствует о том, что самоконтролируемые Vision Transformers эффективно улавливают долгосрочные зависимости и объемную непрерывность. Данное исследование подчеркивает потенциал самоконтролируемых обучающих фреймворков на основе Vision Transformers для преодоления ограничений традиционных методов глубокого обучения с учителем. Предложенный метод продемонстрировал, что он может снизить зависимость от больших полностью выборочных наборов данных и поддерживать точные трехмерные реконструкции для клинических и исследовательских приложений.

Ключевые слова: коленный сустав, магнитно-резонансная томография, трехмерная визуализация, передняя крестообразная связка, сегментация изображений, трансформеры, искусственный интеллект.

https://doi.org/10.52889/1684-9280-2025-76-6-jto027