Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Demonstrations Track

TRIKOP: Exploring Visual Prompting Paradigms for Multi-Grade Knee Osteoarthritis Classification on MRI Images

Hieu Phan[1,2], **Hung Pham**[1,2], **Dat Nguyen**[4], **Khoa Le**[2,3], **Tuan Nguyen**[4], **Triet Tran**[2,3] and **Tho Quan**[1,2,*]
[1]Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam
[3] Ho Chi Minh City University of Science (HCMUS), Ho Chi Minh City, Vietnam
[4]Global Softwares Corporation (GSOFT CORPORATION), Ho Chi Minh City, Vietnam
{hieupt, phhung.sdh241, qttho}@hcmut.edu.vn, {datnt, tuannt}@gsoft.com.vn,
lpmkhoa22@apcs.fitus.edu.vn, tmtriet@fit.hcmus.edu.vn,

Abstract

Knee osteoarthritis (KOA) is a degenerative joint disease that significantly impacts quality of life. While transfer learning shows promise in medical imaging, its application to KOA diagnosis remains challenging due to medical data's unique characteristics. To address this, we propose TRIKOP, a framework leveraging Visual Prompting for KOA diagnosis on MRI. Our approach explores three prompt-generating strategies that extract clinically relevant information from input images. Each prompt type is encoded using a tailored method to integrate effectively into the Vision Transformer for optimal representation. Among them, the contrastive embedding prompting strategy achieves 63.04% accuracy on the OAI dataset, surpassing prior studies. Moreover, TRIKOP produces attention maps highlighting diagnostically significant regions, improving model interpretability. This work highlights TRIKOP's potential to improve AI-driven KOA diagnosis and clinical support.

1 Introduction

Knee osteoarthritis (KOA) is a prevalent degenerative condition affecting the knee joint and is the most common form of osteoarthritis [Prieto-Alhambra *et al.*, 2014], significantly reducing quality of life. The rapid advancement of Artificial Intelligence (AI) in healthcare has unlocked immense potential for analyzing medical imaging data and developing precise diagnostic systems. Vision Transformers (ViTs) [Dosovitskiy *et al.*, 2021], renowned for their ability to capture global image relationships through self-attention [Vaswani *et al.*, 2017], excel in computer vision tasks when trained on large datasets. However, their application in medical imaging remains challenging due to the scarcity of labeled data and the distinct characteristics of medical images. Transfer learning with pretrained ViT models, particularly through prompting methods like Visual Prompt Tuning (VPT) [Jia *et al.*, 2022], has emerged as an effective adaptation approach. VPT employs learnable soft prompts for task-specific tuning, but its reliance on large datasets can limit its applicability in the medical domain, where data availability is often constrained.

To address these limitations, we propose using domain-specific medical knowledge to design hard visual prompts that better guide pretrained ViTs. Rather than learning prompts from limited data, our method embeds meaningful priors to improve performance. We introduce TRIKOP—a trilogy of visual prompting strategies combining anomaly detection, GradCAM heatmaps, and contrastive learning—for KOA severity classification on MRI with frozen ViTs.

Our contributions are as follows. We propose TRIKOP with three novel hard visual prompting paradigms designed specifically for KOA severity classification, leveraging domain-specific medical knowledge to guide pretrained ViT effectively. Additionally, we develop a web application[1] [2] that allows users to upload MRI scans, evaluate the proposed prompting methods, and visualize their outputs.

2 Related Work

2.1 Knee Osteoarthritis Diagnosis

Numerous studies have applied AI to diagnose KOA using medical imaging. For X-ray images, [Antony *et al.*, 2017] employed CNNs and [Tiulpin *et al.*, 2018] leveraged Siamese networks to classify disease severity. Transformer-based models have also been explored, with [Wang *et al.*, 2023] applying ViT and modifying the original Positional Embedding of ViT to enhance feature extraction and improve performance.

MRI, offering richer structural details, has increasingly been adopted. [Guida *et al.*, 2021] demonstrated that 3D CNNs outperform 2D X-rays in assessing severity. [Schiratti *et al.*, 2021] used EfficientNet-B0 [Tan and Le, 2019] and Attention mechanisms to weigh the importance of MRI slices. [Berrimi *et al.*, 2024] proposed a 3D CNN combined with Global Average Pooling and Residual Connections, improving feature extraction and prediction accuracy.

*Corresponding author

[1]https://gmedai.com/app/main/mri-diagnosis
[2]https://github.com/GlobalAILab/TRIKOP

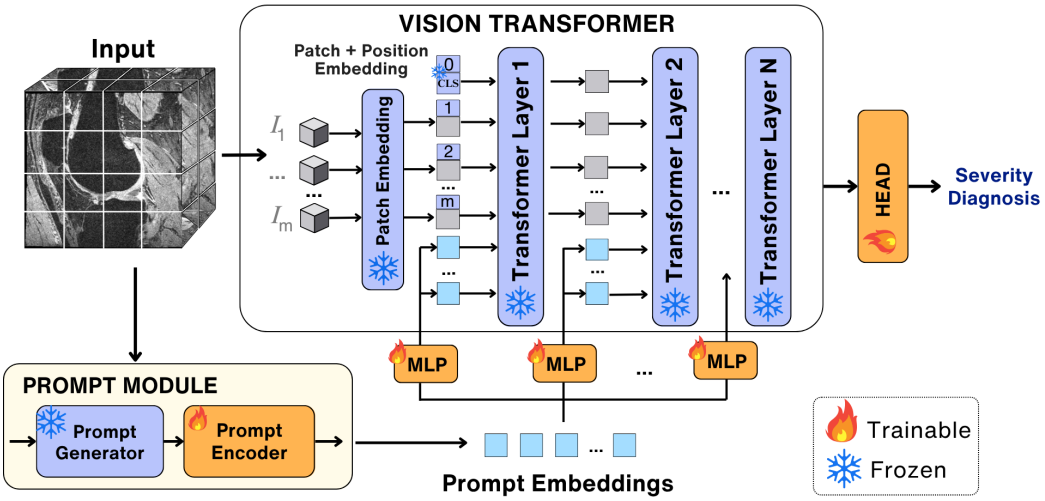

Figure 1: Overview of the TRIKOP framework architecture. The core component, the Prompt Module, generates medical context-aware prompts to effectively adapt a pre-trained large Vision Transformer for knee osteoarthritis severity classification.

2.2 Visual Prompting

The use of large pretrained models on extensive datasets for specific tasks is increasingly popular. Inspired by text prompting in NLP, [Bahng *et al.*, 2022] and [Jia *et al.*, 2022] explored visual prompting for vision models, demonstrating its potential in vision tasks. This approaches has since paved the way for promising applications in medical imaging. Medical Visual Prompting (MVP) [Chen *et al.*, 2024] utilizes patch embeddings and superpixels for disease-specific segmentation, while DVPT [He *et al.*, 2025] extracts sample and domain-specific features to reduce domain gaps with minimal trainable parameters. Similarly, [Denner *et al.*, 2025] employs visual prompt engineering to enhance BiomedCLIP [Zhang *et al.*, 2025], embedding visual cues in radiographs to improve focus on critical regions.

3 The TRIKOP Framework

In this section, we introduce the TRIKOP framework, which leverages visual prompting to efficiently and effectively adapt pretrained vision transformers for KOA diagnosis.

3.1 Overview Methodology

The overall framework of TRIKOP is illustrated in Fig. 1, consisting of three main components: a frozen ViT backbone, a prompt module to generate prompt tokens, and a classification head to predict the severity of KOA.

The input 3D MRI volume $I \in \mathbb{R}^{D\times H\times W}$ is divided into m fixed-sized patches $\{I_j \in \mathbb{R}^{D'\times H'\times W'} \mid j = 1..m\}$ and embedded into a d-dimensional latent space with positional encoding, as in the original ViT. Simultaneously, input I is passed through a prompt module consisting of a prompt generator to create medical-context-aware prompt based on the input I and a prompt encoder to encode them into embeddings $P = \{P^j \in \mathbb{R}^d \mid j = 1, 2, \ldots, p\}$, where p is the number of generated embeddings. These embeddings are fed into transformer layers along with image patch embeddings and the classification token ($[CLS]$)'s embedding, guiding the model's attention to relevant features. This process can be expressed as:

$$[x_i, \_, E_i] = L_i([x_{i-1}, P_{i-1}, E_{i-1}]), \quad i = 1..N \qquad (1)$$

where L_i is the i-th Transformer layer, $x_{i-1} \in \mathbb{R}^d$ is the $[CLS]$ token embedding, $P_{i-1} \in \mathbb{R}^{p\times d}$ represents the prompt embeddings after passing P through the MLP_{i-1}, and $E_{i-1} \in \mathbb{R}^{m\times d}$ is the image patch embeddings. After N transformer layers, the updated $[CLS]$ token x_N is passed through a fully connected layer for final KOA severity prediction.

To construct effective prompt embeddings P, we propose three paradigms for the prompt module: the anomaly map, the Grad-CAM heatmap, and the contrastive embedding paradigm.

3.2 Paradigm 1: Anomaly Map-based Prompting

In this paradigm, we leverage the Osteo-GAN model from our previous study [Phan *et al.*, 2024] to generate anomaly maps using a reconstruction-based technique. Specifically, we pretrained an Osteo-GAN model on 50-th slices extracted from MRIs, as the 50-th slice represents the central region and generalizes well across knee MRI scans. This pretrained Osteo-GAN model serves as the prompt generator, as illustrated in Fig. 2.

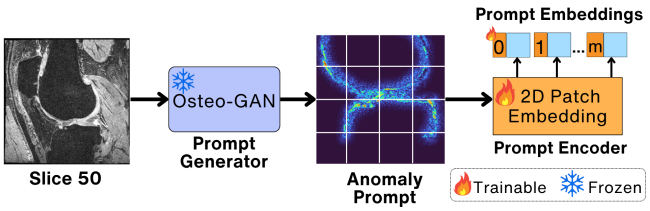

Figure 2: Architecture of the Anomaly Prompt module.

Given a 3D input image $I \in \mathbb{R}^{D\times H\times W}$, the prompt generator G_{Ano} produces an anomaly map for the 50-th slice $M_{\text{Ano}} = G_{\text{Ano}}(I_{50}) \in \mathbb{R}^{H\times W}$.

The anomaly map M_{Ano} is then passed through the prompt encoder $\text{Enc}_{\text{Ano}}(\cdot)$, which applies 2D patch embedding with positional encoding, similar to the embedding mechanism in ViT, to generate prompt embeddings $P = \text{Enc}_{\text{Ano}}(M_{\text{Ano}}) \in \mathbb{R}^{p \times d}$.

3.3 Paradigm 2: GradCam Heatmap-based Prompting

Anomaly-based prompting, as a 2D prompt applied to 3D medical images, can introduce noise due to limited spatial representation. To address this, we propose Paradigm 2, which uses GradCAM [Selvaraju *et al.*, 2017] to generate 3D heatmaps from a 3D CNN model. These heatmaps highlight key regions influencing the model's predictions, guiding the ViT to focus on medically relevant areas and improving classification performance. Fig. 3 illustrates this paradigm, which uses GradCAM heatmaps as prompts.

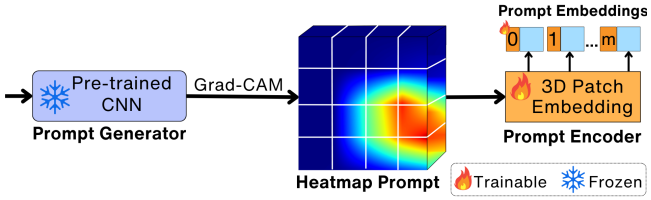

Figure 3: Architecture of the GradCAM Heatmap Prompt module.

Firstly, we pretrained a supervised simple 3D CNN on the MRI dataset with a three-classification task (healthy (KL-0), early-stage (KL-1 & KL-2), advanced (KL-3 & KL-4)). This pretrained 3D CNN model is then employed as the generator G_{CAM} in the prompt module. Given a 3D input image $I \in \mathbb{R}^{D \times H \times W}$, the prompt generator G_{CAM} generates a GradCAM-based heatmap $M_{\text{CAM}} = G_{\text{CAM}}(I) \in \mathbb{R}^{D \times H \times W}$.

Subsequently, the prompt encoder $\text{Enc}_{\text{CAM}}(\cdot)$, which applies 3D patch embedding followed by positional encoding (similar to the embedding layer of ViT), encodes the heatmap into prompt embeddings $P = \text{Enc}_{\text{CAM}}(M_{\text{CAM}}) \in \mathbb{R}^{p \times d}$ in the TRIKOP framework.

3.4 Paradigm 3: Contrastive Learning-based Prompting

While Grad-CAM-based prompting enhances performance with detailed spatial information, it doubles the token count, significantly increasing self-attention costs in the transformer. To address this, Paradigm 3 reduces token complexity while retaining essential feature representations. Fig. 4 illustrates this paradigm, which employs contrastive features as prompts.

We first pretrain a 3D CNN model, referred to as the prompt generator G_{SupCon}, using supervised contrastive learning [Khosla *et al.*, 2020] to capture representations of disease severity for prompt generation.

The input image $I \in \mathbb{R}^{D \times H \times W}$ is first processed by G_{SupCon}, yielding a feature map $M_{\text{SupCon}} = G_{\text{SupCon}}(I) \in \mathbb{R}^{\frac{D}{15} \times \frac{H}{32} \times \frac{W}{32}}$, which encodes key discriminative features. This feature map is then flattened into a contrastive embedding P_{SupCon}. Finally, P_{SupCon} is projected into the d-dimensional space by a linear projection layer $\text{Enc}_{\text{SupCon}}(\cdot)$, generating the final prompt embeddings $P = \text{Enc}_{\text{SupCon}}(P_{\text{SupCon}}) \in \mathbb{R}^{p \times d}$.

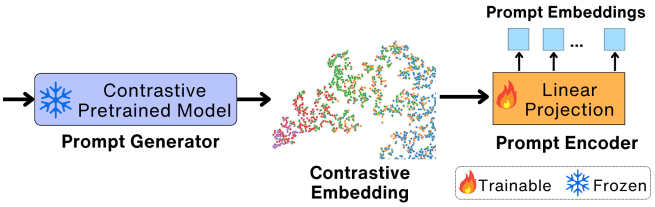

Figure 4: Architecture of the Contrastive Embedding Prompt module.

4 Experiment and Results

We utilized 3D DESS MRI sequences from the Osteoarthritis Initiative (OAI) project [National Institute of Mental Health, 2001]. Following the preprocessing pipeline proposed by [Guida *et al.*, 2021], we modified the MRI volumes from their original resolution of $160 \times 384 \times 384$ to $120 \times 160 \times 160$ by cropping and removing non-informative slices. Instead of using center cropping as in [Guida *et al.*, 2021], we applied automated cropping based on segmentation results to precisely capture the region containing the cartilage between the two bones.

Method	Accuracy (%)	#Params
Baseline		
[Guida *et al.*, 2021]	54.00†	-
Full fine-tune	61.09	88.2
Visual Prompting		
VPT-Deep [Jia *et al.*, 2022]	53.26	0.06
TRIKOP Anomaly Map	59.29	28.6
TRIKOP GradCAM Heatmap	61.61	31.5
TRIKOP Contrastive Embedding	63.04	29.9

#Params refers to the number of trainable parameters (Million).
† Results are taken from [Guida *et al.*, 2021]

Table 1: Comparison of results between prompting methods and the baseline.

Table 1 presents the benchmark results of three paradigms compared to other approaches. The results demonstrate that our visual prompting method achieves competitive accuracy with full fine-tuning while requiring significantly fewer trainable parameters.

5 Conclusion

In this study, we introduced TRIKOP, an application that utilizes Visual Prompting to enhance the diagnostic outcomes of KOA. It also enables users to visualize the affected areas on the knee joint that the model focuses on to produce diagnostic results through attention maps. In the future, we plan to refine prompt methods to improve the model's diagnostic accuracy. We anticipate that this application will bring significant value to patients and doctors in the treatment of KOA.

Acknowledgments

This research is funded by Viet Nam National University Ho Chi Minh City (VNU-HCM) under grant number DS-2025-20-05. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

We sincerely thank Prof. Truyen The Tran, Head of AI, Health & Science at Deakin University, for his insightful suggestion on applying visual prompting to KOA diagnosis. His valuable input not only inspired this research but also provided constructive feedback that further refined both the study and manuscript.

References

[Antony *et al.*, 2017] Joseph Antony, Kevin McGuinness, Kieran Moran, and Noel E. O'Connor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 376–390, Cham, 2017. Springer International Publishing.

[Bahng *et al.*, 2022] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

[Berrimi *et al.*, 2024] Mohamed Berrimi, Didier Hans, and Rachid Jennane. A semi-supervised multiview-mri network for the detection of knee osteoarthritis. *Computerized Medical Imaging and Graphics*, 114:102371, 2024.

[Chen *et al.*, 2024] Yulin Chen, Guoheng Huang, Kai Huang, Zijin Lin, Guo Zhong, Shenghong Luo, Jie Deng, and Jian Zhou. Medical visual prompting (mvp): A unified framework for versatile and high-quality medical image segmentation, 2024.

[Denner *et al.*, 2025] Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, and Klaus Maier-Hein. Visual prompt engineering for vision language models in radiology, 2025.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[Guida *et al.*, 2021] Carmine Guida, Ming Zhang, and Juan Shan. Knee osteoarthritis classification using 3d cnn and mri. *Applied Sciences*, 11(11), 2021.

[He *et al.*, 2025] Along He, Kai Wang, Zhihong Wang, Tao Li, and Huazhu Fu. Dvpt: Dynamic visual prompt tuning of large pre-trained models for medical image analysis. *Neural Networks*, 185:107168, 2025.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[National Institute of Mental Health, 2001] National Institute of Mental Health. The Osteoarthritis Initiative. https://nda.nih.gov/oai, 2001. Accessed on: February 2024.

[Phan *et al.*, 2024] Hieu Phan, Loc Le, Mao Nguyen, Phat Nguyen, Sang Nguyen, Triet Tran, and Tho Quan. Xgaosteo: towards xai-enabled knee osteoarthritis diagnosis with adversarial learning. In *IJCAI*, 2024.

[Prieto-Alhambra *et al.*, 2014] Daniel Prieto-Alhambra, Andrew Judge, Kassim M. Javaid, Cyrus Cooper, Adolfo Diez-Perez, and Nigel K. Arden. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. *Annals of the Rheumatic Diseases*, 73(9):1659–1664, 2014.

[Schiratti *et al.*, 2021] Jean-Baptiste Schiratti, Rémy Dubois, Paul Herent, David Cahané, Jocelyn Dachary, Thomas Clozel, Gilles Wainrib, Florence Keime-Guibert, Agnes Lalande, Maria Pueyo, Romain Guillier, Christine Gabarroca, and Philippe Moingeon. A deep learning method for predicting knee osteoarthritis radiographic progression from mri. *Arthritis Research & Therapy*, 23(1):262, 2021.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97, pages 6105–6114. PMLR, 09–15 Jun 2019.

[Tiulpin *et al.*, 2018] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Scientific reports*, 8(1):1727, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

[Wang *et al.*, 2023] Zhe Wang, Aladine Chetouani, Mohamed Jarraya, Didier Hans, and Rachid Jennane. Transformer with selective shuffled position embedding using roi-exchange strategy for early detection of knee osteoarthritis. *arXiv preprint arXiv:2304.08364*, 2023.

[Zhang *et al.*, 2025] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn,

Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025.