한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 4, P. 47~56

LoRA 기반 멀티모달 파운데이션 모델 적응을 통한 MR 영상 반월상 연골 분할

조은빈[1O] 이민진[2O] 왕준호[3] 홍헬렌[2*]

[1]서울여자대학교 데이터사이언스학과
[2]서울여자대학교 소프트웨어학과
[3]성균관대학교 의과대학 삼성서울병원 정형외과
(binning, minjin, hlhong)@swu.ac.kr, mdwang88@gmail.com

LoRA-Based Multimodal Foundation Model Adaptation for MR Meniscus Segmentation

Eunbin Cho[1O] Min Jin Lee[2O] Joon Ho Wang[3] Helen Hong[2*]

[1] Department of DataScience, Seoul Women's University
[2]Department of Computer Science, Seoul Women's University
[3]Department of Orthopedic Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine

요 약

본 논문에서는 대규모 복부 CT 데이터로 사전 학습된 멀티모달 파운데이션 모델과 매개변수 효율적 미세조정 기법인 LoRA를 결합하여, 데이터가 제한적인 환경에서도 무릎 MR 영상 내 반월상 연골을 안정적으로 분할할 수 있는 방법을 제안한다. 제안 방법은 영상 정보와 해부학적 의미를 담은 텍스트 프롬프트를 통합적으로 활용하고, 이미지 및 텍스트 모듈의 레이어에 LoRA를 적용함으로써 연산 효율성과 분할 성능을 동시에 향상시킨다. 실험 결과, 제안 모델은 전체 파라미터의 일부만을 학습에 활용하면서도 외측 및 내측 반월상 연골 모두에서 높은 분할 성능을 달성하였으며, 특히 소량의 학습 데이터 환경에서도 안정적인 성능을 유지함을 확인하였다. 또한 LoRA의 적용 위치에 따른 분할 성능 변화를 체계적으로 분석하여, 반월상 연골 분할에 적합한 효율적인 LoRA 적용 구조를 제시하였다.

Abstract

In this paper, we propose a method for stable meniscus segmentation in knee MR images under data-limited settings by combining a multimodal foundation model pre-trained on large-scale abdominal CT data with a parameter-efficient fine-tuning technique, LoRA. The proposed method jointly exploits image information and text prompts encoding anatomical semantics, and applies LoRA to selected layers of the image and text modules to improve both computational efficiency and segmentation performance. Experimental results demonstrate that the proposed model achieves high segmentation accuracy for both lateral and medial menisci while updating only a subset of the total parameters, and maintains stable performance even with a limited number of training samples. In addition, by systematically analyzing the segmentation performance according to the placement of LoRA modules, we identify an efficient LoRA adaptation strategy tailored for meniscus segmentation.

[O] These authors contributed equally to this work.
[*] corresponding author

*corresponding author: Helen Hong / Department of Computer Science, Seoul Women's University (hlhong@swu.ac.kr)

Received : 2026.02.04./ Review completed : 1st 2026.04.24. 2nd 2026.06.08./ Accepted : 2026.06.16.
DOI : 10.15701/kcgs.2026.32.4.47
ISSN : 1975-7883(Print)/2383-529X(Online)

키워드: 반월상 연골 분할, 자기공명영상, 멀티모달 파운데이션 모델, 저랭크 적응, 매개변수 효율적 미세조정

Keywords: Meniscus Segmentation, Magnetic Resonance Imaging (MRI), Multimodal Foundation Model, Low-Rank Adaptation (LoRA), Parameter-Efficient Fine-Tuning

1. 서론

반월상 연골은 무릎 안쪽의 내측 반월상 연골과 바깥쪽의 외측 반월상 연골로 구성된 반달 모양의 구조물로, 무릎의 하중을 분산시키고 대퇴골과 경골 사이의 마찰을 줄여 관절을 보호하는 핵심적인 역할을 수행한다[1]. 이러한 반월상 연골의 손상은 무릎의 생체역학적 기능을 저하시켜 조기 골관절염을 유발할 수 있어 적절한 진단과 치료가 필수적이다. 반월상 연골이 심각하게 손상되거나 파열된 경우 반월상연골 이식술(MAT: Meniscal Allograft Transplantation)이 시행되는데, 이때 환자의 반대쪽 정상 무릎을 미러링하여 환자에게 가장 적합한 크기와 형태의 이식편을 제작하는 방법이 주로 사용된다[2]. 따라서 성공적인 이식 수술을 위해서는 반월상 연골의 길이, 너비, 높이 등 3차원적 구조 정보를 정확히 파악할 수 있도록 반대쪽 정상 무릎의 반월상 연골을 자동으로 분할하는 기술이 필요하다.

그림 1은 무릎 MR 영상에서 나타나는 반월상 연골의 특징을 나타낸다. 반월상 연골은 십자인대나 측부 인대와 같은 주변 연부조직과 매우 유사한 밝기값을 가지므로 주변 조직 간의 경계가 명확하게 구분하기 어려운 한계가 있으며, 반월상 연골 내부에서도 신호 강도의 불균일성이 나타나 단일 구조물임에도 불구하고 내부 텍스처가 일관되지 않게 표현되는 경우가 빈번하다. 또한, 반월상 연골은 동일 환자의 영상 내에서도 슬라이스 위치에 따라 전각(anterior horn), 중체(mid body), 후각(posterior horn)의 단면 형태가 급격하게 변화하며, 전각과 후각 부위에서는 크기가 작고 경계의 불명확성이 두드러지게 나타나기 때문에, 이로 인해 분할을 자동화하기에 어려움이 있다[3].

무릎 MR 영상에서 반월상 연골 분할은 주로 지도학습 기반의 합성곱 신경망(CNN)을 통해 수행되어 왔다. 특히 의료 영상 분할의 가장 일반적으로 사용되는 U-Net[4] 구조가 도입된 이후, 이를 반월상 연골 분할 연구에 적용하거나 개선하려는 다양한 시도가 활발히 이루어졌다. Norman 등[5]은 반월상 연골의 정량적 이완 시간과 형태학적 특성을 효율적으로 추출하기 위해 2D-UNet 기반의 분할 방법을 제안하였다. Jeon 등[6]은 반월상 연골이 배경 영역에 비해 차지하는 비중이 작아 발생하는 클래스 불균형 문제를 해결하기 위해 2D-UNet으로 관심 영역(ROI: Region Of Interest, ROI)을 추출하고, 이후 적대적 학습과 객체 인식 맵을 통해 과소 및 과대 분할 오류를 보정하도록 하는 반복적 개선 방법을 적용하여 2단계 심층 합성곱 신경망(DCNN)을 제안하였다. 그러나 이러한 지도 학습 기반 방법론들은 높은 성능을 달성하기 위해서는 픽셀 단위의 정교한 대규모 데이터셋이 필요하며, 레이블 데이터셋이 부족한 경우 성능 저하를 발생시킬 수 있는 한계가 있다.

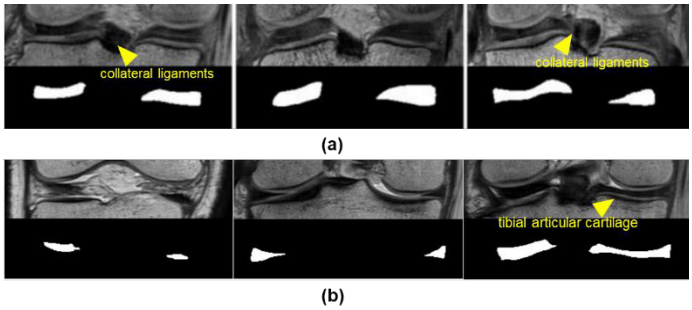

Figure 1. Characteristic of meniscus in knee MR images: (a) similar intensity of meniscus and collateral ligaments or tibial articular cartilage, (b) Diverse shape and thin structure of the meniscus across anterior horn, mid body, and posterior horn slices in a single patient's coronal view

최근 딥러닝 연구의 패러다임은 특정 작업에 특화된 모델을 처음부터 학습시키는 방식에서 벗어나, 대규모 데이터로 사전 학습된 파운데이션 모델(foundation model)을 미세조정(fine-tuning)하여 활용하는 방향으로 전환되고 있다. 특히 자연 영상 분야에서 우수한 제로샷(Zero-Shot) 분할 성능을 입증한 SAM(Segment Anything Model)[7]이 등장함에 따라, 이를 의료 영상 분석에 접목하려는 시도가 급증하고 있다. Mazurowski 등[8]은 19종의 공개 의료 데이터셋을 통해 SAM의 성능을 광범위하게 평가하였으며, 간이나 신장처럼 주변 조직과의 대조도가 비교적 명확한 복부 CT 영상에서 별도의 추가 학습 없이도 준수한 분할 성능을 보임을 입증하였다. 그러나 이러한 기존 SAM 방식은 추론 과정에서 점이나 박스와 같은 기하학적 프롬프트를 사용자가 입력해 주어야 한다는 한계가 있다. 특히, 입력되는 프롬프트의

위치나 개수, 박스의 정밀도에 따라 분할 성능이 크게 달라지는 민감도 문제가 존재하여 일관된 성능을 보장하는 완전 자동화된 의료 영상 분석 모델로 적용하기에는 어려움이 따른다. 또한 의료 영상 분할 분야에서는 대규모 데이터를 활용하여 다양한 장기와 종양을 범용적으로 분할할 수 있는 유니버설 모델에 대한 연구가 진행되고 있으며, Liu등[9]은 자연어 처리 분야에서 뛰어난 성능을 보인 CLIP(Contrastive Language-Image Pre-training)을 의료 영상 분할에 접목한 CLIP-Driven Universal Model을 제안하였다. 이 모델은 기존의 원-핫 인코딩(One-hot encoding) 방식이 장기 간의 의미적 연관성을 무시한다는 한계를 극복하기 위해, 텍스트 임베딩을 활용하여 해부학적 구조의 의미적 관계를 학습하였다. 또한, 14개의 공개 데이터셋을 통합한 3,410건의 대규모 CT 스캔을 통해 학습되었으며, MSD 및 BTCV와 같은 벤치마크에서 SOTA 성능을 달성하여 텍스트 기반 유도(Text-driven guidance)의 유효성을 입증하였다. 그러나 이 모델은 주로 CT 영상을 기반으로 구축되었기 때문에, 물리적 특성과 신호 강도 분포가 상이한 MR 영상에 직접 적용하기에는 도메인 격차가 존재한다. 또한 파라미터 수가 많은 거대 모델 특성상 새로운 도메인에 적용시키기 위해서는 전체를 미세조정 하는 것은 막대한 계산 비용과 데이터가 요구된다.

따라서 본 논문에서는 CT 기반 파운데이션 모델의 의미적 표현 능력을 유지하면서도, 적은 연산 비용으로 MR 영상의 반월상 연골 분할에 최적화된 멀티모달 파운데이션 모델 적응 프레임워크를 제안한다. 이를 위해 매개변수 효율적 미세조정(parameter efficient adaptation) 기법인 LoRA(Low-Rank Adaptation)[10]를 도입하여, 데이터가 제한적인 환경에서도 모델이 MR 도메인의 특성을 효율적으로 학습할 수 있도록 설계하였다. 또한 영상 정보뿐만 아니라 해부학적 의미를 담은 텍스트 정보를 통합함으로써, 경계가 모호한 영역에서도 분할 성능을 효과적으로 향상시키고자 한다.

본 논문의 주요 기여점은 다음과 같다. 첫째, 높은 연산량과 대규모 학습 데이터를 요구하는 기존의 전체 미세조정 방식을 대체할 수 있는 LoRA기반의 파라미터 효율적 미세조정 방법을 제안한다. 특히 LoRA의 적용 위치에 따른 성능 변화를 체계적으로 분석하여, 반월상 연골 분할에 적합한 효율적 적용 구조를 제시한다. 둘째, 시각적 정보에만 의존하던 기존 분할 모델의 한계를 극복하기 위해 텍스트 임베딩을 통합하는 멀티모달 융합 메커니즘을 도입한다. 이를 통해 영상 내 대조도가 낮고 경계가 불분명한 영역에서도 텍스트 정보를 분할 가이드로 활용할 수 있는 가능성을 제시한다.

2. 제안 방법

2.1 개요

그림 2는 본 논문에서 제안하는 전체 프레임워크의 개요를 나타낸다. 대규모 CT 데이터셋을 통해 사전 학습된 CLIP-Driven Universal Model을 파운데이션 모델로 활용하고, 이를 기반으로 MR 영상에서의 반월상 연골 분할을 수행하기 위해 모델을 적응(adaptation) 한다. 또한, 해부학적 구조 정보를 효과적으로 반영하기 위해 텍스트 인코더를 통해 생성된 구조별 텍스트 임베딩을 영상 특징과 결합하여 분할을 유도한다. 제안 방법은 내측 및 외측 반월상 연골을 각각 독립적인 구조로 모델링함으로써, 서로 다른 해부학적 형태와 위치적 특성을 고려한 분할을 가능하게 한다. 이 과정에서 파라미터 효율적인 학습을 위해, 파운데이션 모델 내 파라미터 수가 큰 MLP(Multi-Layer Perceptron) 기반 모듈에는 LoRA를 적용하여 저차원 적응을 수행한다. 반면 상대적으로 파라미터의 수가 적은 컨볼류션 기반 모듈은 전체 파라미터를 직접 미세조정 하는 방식을 사용하도록 설계한다. 이후 2.2절에서는 본 논문에서 활용한 파운데이션 모델의 구조와 사전 학습 특성을 설명하고, 2.3절에서는 LoRA 기반 적응 전략을 중심으로 파라미터 효율화 기법을 상세히 기술한다.

2.2 이미지-텍스트 멀티모달 유니버설 파운데이션 모델

본 논문에서 사용하는 파운데이션 모델인 CLIP-Driven Universal Model[9]은 이미지와 텍스트 정보를 결합하여 해부학적 구조를 이해하도록 설계된 멀티모달 범용 모델로서, Figure 2와 같이 이미지 모듈, 텍스트 모듈, 그리고 두 정보를 통합하는 텍스트-컨트롤러 모듈의 세 가지 핵심 요소로 구성된다.

이미지 모듈은 3차원 의료 영상으로부터 다중 스케일의 시각적 특징을 추출하고, 이를 입력 영상과 동일한 해상도로 복원하여 정밀한 분할을 위한 고해상도 특징 맵을 생성한다. 본 모듈은 비전 트랜스포머(ViT) 계열인 Swin Transformer[11]기반의 인코더와 3차원 합성곱 신경망(CNN) 기반의 디코더가 결합된 SwinUNETR를 백본으로 사용한다. 해당 구조는 이동 윈도우(Shifted Window)기반의 자기 어텐션(Self attention)을 통해 전역적 문맥 정보를 학습하는

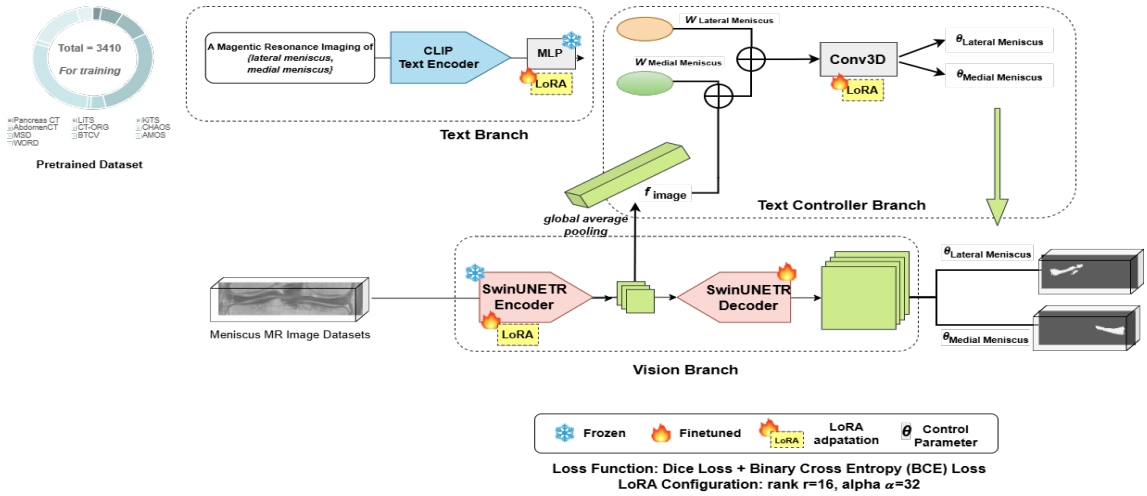

Figure 2. Overview of proposed CLIP-Driven Universal Model

트랜스포머의 장점과 국소적인 공간적 정보를 효과적으로 포착하는 CNN의 표현력을 동시에 활용할 수 있다는 장점이 있다. 인코더는 총 4개의 단계로 구성되며, 각 단계는 2개의 SwinTransformer 블록을 포함하고 있어 총 8개의 레이어로 구성된다. 각 단계의 끝단에는 패치 병합(Patch Merging) 계층이 적용되어, 해상도를 2배씩 감소시키며 특징을 압축한다. 디코더는 전치 합성곱(transposed convolution) 연산을 통해 업샘플링(upsampling)한 뒤, 인코더의 스킵 연결을 통해 전달된 특징과 결합하여 잔차 블록을 통과시키는 과정을 반복함으로써 고수준의 의미론적 정보와 공간적 세부 정보를 효과적으로 복원한다.

텍스트 모듈은 자연어로 표현된 분할 대상의 개념적 의미를 정량적인 벡터 형태로 변환하여, 시각적 특징과 결합 가능한 의미 기반 조건을 제공한다. 이를 위해 사전 학습된 CLIP 텍스트 인코더를 사용하며, 본 연구의 대상인 반월상 연골을 지칭하기 위해 사용하며 "A Magnetic Resonance Imaging of {lateral meniscus, medial meniscus}" 와 같은 프롬프트를 입력으로 사용한다. CLIP 텍스트 인코더는 대규모 이미지-텍스트 쌍을 통해 학습된 사전 지식을 바탕으로, 입력 텍스트를 해당 해부학적 구조의 추상적 개념과 시각적 맥락을 함축하는 고차원 임베딩으로 변환한다.

텍스트-컨트롤러 모듈은 이미지 모듈에서 추출된 전역적 시각 특징과 텍스트 임베딩을 결합하여, 특정 장기 분할에 최적화된 동적 파라미터를 생성한다. 이미지 특징은 전역 평균 풀링을 통해 벡터 형태로 변환될 후 텍스트 임베딩과 채널 방향으로 결합되어 컨트롤러에 입력한다. 컨트롤러는 입력된 시각-언어 정보를 기반으로 3차원 합성곱 연산을 통해 특정 장기별 동적 파라미터를 생성하며, 이는 디코더 마지막 단계에서 위치한 텍스트-구동 분할기(Text-Driven Segmentor) 내의 1x1x1 합성곱 레이어의 가중치로 적용한다. 이로써 모델은 고해상도 특징 맵에서 텍스트로 지정된 반월상 연골 구조에 선택적으로 집중하여 최종 이진 분할 마스크를 생성한다.

2.3 LoRA 적응을 통한 매개변수 효율화

본 논문에서는 대규모 복부 CT 데이터셋으로 사전 학습된 CLIP-Driven Universal Model을 파운데이션 모델로 채택하였으나, 이를 별도의 미세조정 없이 MR 영상에 직접 적용할 경우 유의미한 성능 저하가 발생할 수 있다. 이는 CT와 MR 영상 간의 모달리티 격차(Modality Gap)와 복부 장기와 반월상 연골 간의 해부학적 구조 차이에 따른 도메인 시프트(Domain Shift)에 기인한다. 이러한 문제를 해결하기 위해 모델 전체를 미세조정하는 방식은 막대한 연산 자원을 요구할 뿐만 아니라, 제한된 MR 영상 학습 데이터로 인해 과적합의 위험이 크다. 이에 본 논문에서는 소수의 파라미터만을 갱신하여 모델을 효과적으로 적응시키는 매개변수 효율적 미세조정 기법인 LoRA를 도입하고자 한다.

LoRA는 대규모 사전학습 모델을 특정 작업에 맞게 적응시키는 과정에서 발생하는 계산 및 저장 비용을 줄이기 위해 제안된 기법으로, 사전 학습된 가중치 행렬 W_o 에 새롭게 학습 가능한 저랭크 행렬 A와 B 를 추가하여 가중치 보정값 $\Delta W = BA$ 를 학습한다. 이는 아래 수식 (1) 과 같이 표현할 수 있다.

$$W = W_o + \Delta W, \qquad \Delta W = BA$$

$$B \in R^{d\times r}, A \in R^{r\times k}, r \ll \min(d,k) \quad (1)$$

이때, d 과 k 는 원본 가중치 행렬의 차원이며, r 은 LoRA의 랭크로 설정된 작은 값을 의미한다. 이러한 설정은 충분한 표현력을 제공하면서도 학습해야 할 파라미터의 수를 효과적으로 제한하여 계산 효율성과 메모리 사용량을 동시에 절감할 수 있게 한다. LoRA에서는 초기 행렬 A를 랜덤한 가우시안 분포로 초기화 하고, B는 0으로 초기화함으로써 학습 초기에는 $\Delta W = 0$ 이 되도록 설정하며, 학습 과정에서 사전학습 가중치 W_o 는 고정되고, 오직 A, B 만 최적화된다.

본 논문에서는 LoRA 기법을 CLIP-Driven Universal Model의 모듈 특성에 맞추어 유연하게 적용한다. 일반적인 LoRA가 적용되는 어텐션 연산뿐만 아니라, 텍스트 모듈의 완전 연결 계층과 텍스트-컨트롤러 모듈의 합성곱 레이어까지 적용 범위를 확장하였다. 구체적으로 이미지 모듈의 각 SwinTransformer 블록 내 다중 헤드 자기 어텐션(Multi-Head Self Attention, MSA)에서 쿼리(Query)와 키(Key) 투사층에 LoRA를 적용하여 시각적 특징 학습의 효율성을 높였다. 동일한 방식으로 CLIP 텍스트 인코더 끝단의 MLP 모듈에도 LoRA를 적용하여 텍스트 임베딩의 적응력을 향상시켰다. 또한, 시각 및 텍스트 정보가 융합되는 텍스트-컨트롤러의 3D 합성곱(Conv3D) 레이어에는 커널 차원을 고려하여 확장된 ConvLoRA를 적용함으로써, 전체 모델의 연산 효율성과 미세조정 성능을 동시에 확보한다. 한편, 상대적으로 파라미터 수가 적은 이미지 모듈의 디코더는 전체 파라미터를 직접 미세조정하도록 설계함으로써, LoRA 기반 적응과 결합된 혼합 적응 방법을 사용한다

3. 실험 및 결과

본 논문에서 사용한 데이터는 삼성서울병원으로부터 획득한 총 103개의 무릎 MR 영상으로 구성되며, 기관생명윤리위원회의 승인을 받아 수행되었다. 모든 영상은 Philips Medical Systems의 Achieva 3.0T 장비를 이용해 획득한 3차원 PD VISTA coronal knee MR 영상으로, 해상도는 512 x 512이며 슬라이스 장수는 230-250장, 슬라이스 두께는 0.5mm, 슬라이스 간격은 0.3125mm이다. 데이테셋은 오른쪽 무릎 영상 60개와 왼쪽 무릎 영상 43개로 구성되어 있으며, 좌우 방향에 따른 구조적 차이를 제거하기 위해 왼쪽 무릎 영상은 좌우 반전을 통해 오른쪽 무릎과 동일한 방향으로 정렬하였다. 전체 영상에서 매우 적은 비율을 차지하는 반월상 연골 영역에 집중하기 위해 해부학적 구조를 반영한 ROI를 설정하였다. 전후(front-back) 범위는 대퇴골과 경골이 동시에 나타나는 첫 번째 단면을 기준으로 정의하였으며, 좌우(left-right) 범위는 대퇴골이 처음 나타나는 위치를 기준으로 설정하였다. 상하 범위(top-bottom) 범위는 대퇴골과 경골이 맞닿는 중심선을 기준으로 반월상 연골이 포함될 수 있도록 조정하여 최종 ROI를 결정한다. 이후 해당 ROI를 기준으로 영상을 절단하여 각 슬라이스를 291x80 크기로 정규화하였다. 또한 MR 영상은 촬영 조건 및 환자 특성에 따라 신호 강도의 절대적인 스케일이 일관되지 않으므로, 영상 간 신호 분포 차이를 보정하기 위해 Z-score 정규화를 적용하였다.

모든 실험은 NVIDIA Tesla T4 GPU 환경에서 Python 3.7 기반의 conda 가상환경을 이용하여 수행되었다. 네트워크 학습 시 에폭 수는 40으로 설정하였으며, 학습률은 0.001로 고정하였다. 최적화 알고리즘으로는 Adam 옵티마이저를 사용하였다. 손실 함수로는 다이스(Dice) 손실과 이진 크로스 엔트로피(Binary Cross Entropy Loss) 손실을 함께 사용하였으며, 두 손실 값을 합산하여 최종 손실로 계산하였다. 전체 103개의 데이터 중 학습 데이터로 83개, 테스트 데이터로 20개를 사용하였다. 또한 데이터가 제한적인 환경에서의 학습 효율성을 분석하기 위해, 학습 데이터의 수를 10, 30, 50, 83개로 달리한 추가 실험을 수행하였다. 성능 평가는 다이스 계수, 재현율, 정밀도를 지표로 사용하여 제안 방법과 비교 방법 간의 분할 성능을 평가하였다. 모델 간 성능 차이의 통계적 유의서을 검증하기 위해 대응표본 t-검정(Paried t-test)를 수행하였다. 또한 CLIP 텍스트 인코더 끝단의 MLP에 LoRA 적용 유무에 대한 효과를 검증하기 위하여 t-SNE로 이미지 특징 분포를 확인하였다.

먼저 이미지 모듈의 각 어텐션의 쿼리와 키 투사층에만 LoRA를 적용한 방법을 Model A (Vision encoder + LoRA) 로 정의하였다. 이후 CLIP 텍스트 인코더 끝단의 MLP 모델에 LoRA를 추가 적용한 방법을 Model B(Vision encoder + Text embedding + LoRA)로 구성하였다. 또한, 멀티모달 정보의 효과를 검증하기 위해, CLIP 기반 텍스트 정보를 완전히 배제하고 이미지 모듈로만 구성된 이미지 단일 모달 방식 Model C(Vision-only: Vision encoder + LoRA)를 설정하였다. 이때, Model C는 이미지 인코더에 LoRA를 적용하고 디코더에 전체 미세 조정을 적용한 구조로 구성된다. 본 논문에서 제안하는 방법은 이미지 모듈의 인코더, CLIP 텍스트 인코더와 텍스트-컨트롤러 모두에 LoRA를 적용하고, 이미지 모듈

Table 1. Quantitative comparison of meniscus segmentation performance across model variants with 10 and 83 training samples. Mean and standard deviations are provided, with the highest values highlighted in bold.

Methods	Data		Whole meniscus	Lateral meniscus				Medial meniscus			
	Labled	Unlabeld	DSC(%)	DSC (%)	Recall (%)	Precision (%)	HD95 (pixel)	DSC (%)	Recall (%)	Precision (%)	HD95 (pixel)
2D nnU-Net[12]	22	-	89.40 ±2.0	-	-	-	-	-	-	-	-
	83	-	90.95 ±1.9	-	-	-	-	-	-	-	-
Jeong et al.[3] (Semi-supervised learning)	22	226	90.02 ±2.0	-	-	-	-	-	-	-	-
	83	165	91.11 ±2.3	-	-	-	-	-	-	-	-
Model A	10		0	0	0	0	-	0	0	0	-
	83		5.76 ± 5.77	0	0	0	-	11.51 ±11.54	16.91 ±21.37	10.33 ±8.70	93.55 ±41.39
Model B	10		28.96 ±8.20	23.77 ±8.37	24.55 ±10.42	26.24 ±7.77	35.98 ±44.99	34.15 ±8.02	38.54 ±13.68	32.02 ±7.37	26.18 ±28.44
	83		73.55 ±4.98	69.77 ±5.74	71.31 ±6.19	68.92 ±8.50	6.37 ±3.04	71.78 ±5.51	69.21 ±7.05	75.61 ±8.61	6.59 ±3.89
Model C	10		74.09 ±5.18	72.07 ±8.15	81.38 ±10.26	66.73 ±8.12	17.79 ±30.16	71.63 ±8.62	87.28 ±8.29	60.44 ±10.05	20.07 ±38.66
	83		89.89 ±2.03	89.38 ±2.54	90.83 ±4.24	88.98 ±6.65	**2.25 ±0.57**	90.11 ±2.42	**93.09 ±4.89**	87.61 ±7.14	2.29 ±0.74
Model D	10		86.20 ±3.49	84.87 ±4.77	81.77 ±8.67	89.87 ±6.66	3.58 ±1.21	85.55 ±5.14	79.08 ±9.93	94.22 ±4.84	3.45 ±1.61
	83		**91.17 ±2.37**	**91.49 ±2.68**	**93.71 ±3.94**	**89.80 ±5.94**	2.31 ±0.82	**90.54 ±2.58**	92.31 ±4.20	**89.38 ±6.45**	**2.14 ±0.79**

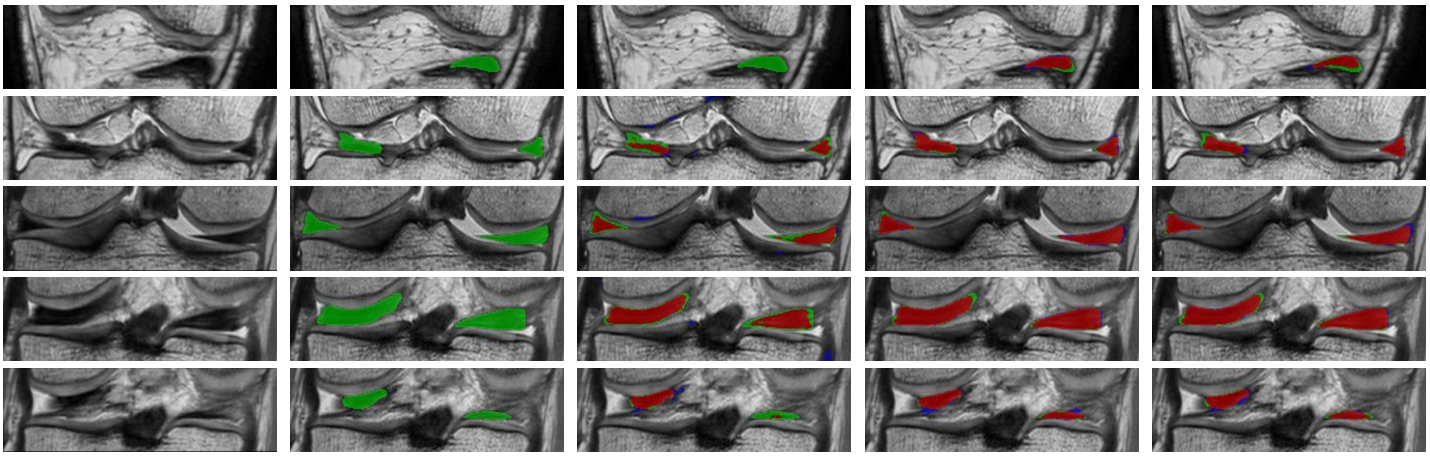

Figure 3. Qualitative comparison of meniscus segmentation results for each model trained with 83 training samples. (Red : overlapping area with ground-truth, Green: under-segmented areas, Blue: over-segmented areas)

디코더에 전체 미세조정을 적용하는 혼합 적응 방법을 사용하여 분할을 수행하는 Model D(Vision encoder + Text embedding + LoRA +Vision decoder fine-tuning, proposed method)이다. 이러한 모델 구성을 통해 제안 방법과 각 비교 모델 간의 분할 성능을 체계적으로 비교 평가하였으며, 모든 비교 실험에서는 LoRA의 랭크 값을 $r = 16$, 스케일링 계수를 $\alpha = 32$로 고정하였다.

표 1은 제안 방법과 비교 방법에 대한 반월상 연골 분할 성능을 정량적으로 비교한 결과를 나타낸다. 또한 제안 방법의 성능을 기존 방법과 비교하기 위해, 동일한 데이터셋 환경에서 평가된 완전지도학습 기반의 2차원 nnUNet과 반지도학습 기반의 Jeong 등의 방법을 함께 제시하였다. 기존

방법들은 외측 및 내측 반월상 연골을 통합한 전체 반월상 연골 기준의 성능을 보고하였으며, 제안 방법의 구성 요소 분석 모델들은 외측 및 내측 반월상 연골에 대한 성능을 각각 평가하였다. 전체 반월상 연골을 기준으로 Model D는 91.17%의 다이스 계수를 기록하여 2D nnU-Net (90.95%) 대비 통계적으로 유의한 성능 향상을 보였다($p < 0.05$). 반면 Jeong 등의 반지도학습 방법과는 유사한 성능을 달성하였으며($p > 0.05$), 추가적인 비라벨 데이터 없이도 동등한 분할 성능을 확보할 수 있음을 확인하였다. Model A는 외측 반월상 연골에 대해 거의 분할 성능을 보이지 않았으며, 내측 반월상 연골에서도 약 11% 수준의 매우 낮은 DSC를 기록하였다. 이는 이미지 모듈의 인코더에 LoRA만을 적용한 경우 CT 영상으로 사전 학습된 분할 모델을 MR 데이터에 적용하는 데 한계가 있음을 시사한다. 반면 Model B는 외측과 내측 반월상 연골에서 각각 69.77%와 71.78%의 다이스 계수를 기록하여 Model A 대비 큰 성능 향상을 보였다. 이는 텍스트 프롬프트를 통해 제공된 해부학적 명칭 정보가 분할 과정에서 의미 있는 조건으로 작용하여 구조의 위치와 형태를 보다 정확하게 인식하도록 유도했기 때문으로 해석된다. 이미지 단일 모달 방식인 Model C는 외측과 내측 반월상 연골에서 평균 약 90% 수준의 다이스 계수를 달성하며 안정적인 분할 성능을 보였다. 제안 방법인 Model D는 Whole meniscus 기준으로 91.17%의 DSC를 기록하여 가장 우수한 성능을 보였다. 또한 외측 및 내측 반월상 연골에서 각각 91.49%와 90.54%의 다이스 계수를 기록하여 Model C 대비 각각 2.11%p와 0.43%p의 성능 향상을 나타냈다. 특히 재현율과 정밀도 측면에서도 균형 잡힌 성능을 보였으며, 이는 텍스트 기반의 해부학적 정보를 분할 조건으로 활용함으로써 반월상 연골에 대한 보다 구별적인 표현을 형성하고, 주변 구조물로의 오분할을 효과적으로 완화할 수 있음을 시사한다. 또한 HD95 기준에서도 Model D는 외측 및 내측 반월상 연골에서 각각 2.31 pixel과 2.14 pixel을 기록하여 Model C와 유사한 수준의 경계 정확도를 보였다. 이는 제안 방법이 영역 중첩 성능(DSC)의 향상뿐만 아니라 구조 경계의 정확성 또한 안정적으로 유지할 수 있음을 시사한다.

그림 3은 83개의 학습 데이터를 사용하여 학습한 각 모델의 반월상 연골 분할 결과를 정성적으로 비교한 결과를 나타낸다. Model A는 대부분의 슬라이스에서 반월상 연골을 적절히 분할하지 못하는 경향을 보였다. Model B는 Model A 대비 분할 성능이 일부 향상되었으나, 전각(anterior horn) 및 후각(posterior horn) 영역에서 분할 누락이 빈번하게 발생하였으며, 중체(mid-body) 영역에서는 경계 부위의 과소 분할이 관찰되었다. 이미지 단일 모달 방식인 Model C는 과소 분할 문제가 상당 부분 완화되어 정답 레이블과 유사한 형태의 분할 결과를 보였으나, 반월상 연골 주변 구조물로의 과대 분할이 발생하는 경향이 확인되었다. 반면 제안 방법인 Model D는 정답 레이블과 가장 높은 일치도를 보였으며, 전각, 중체 및 후각 전 영역에서 반월상 연골을 안정적으로 분할하는 모습을 확인할 수 있었다. 이러한 정성적 결과는 표 1에서 확인된 재현율과 정밀도의 균형적인 향상과도 일관된 경향을

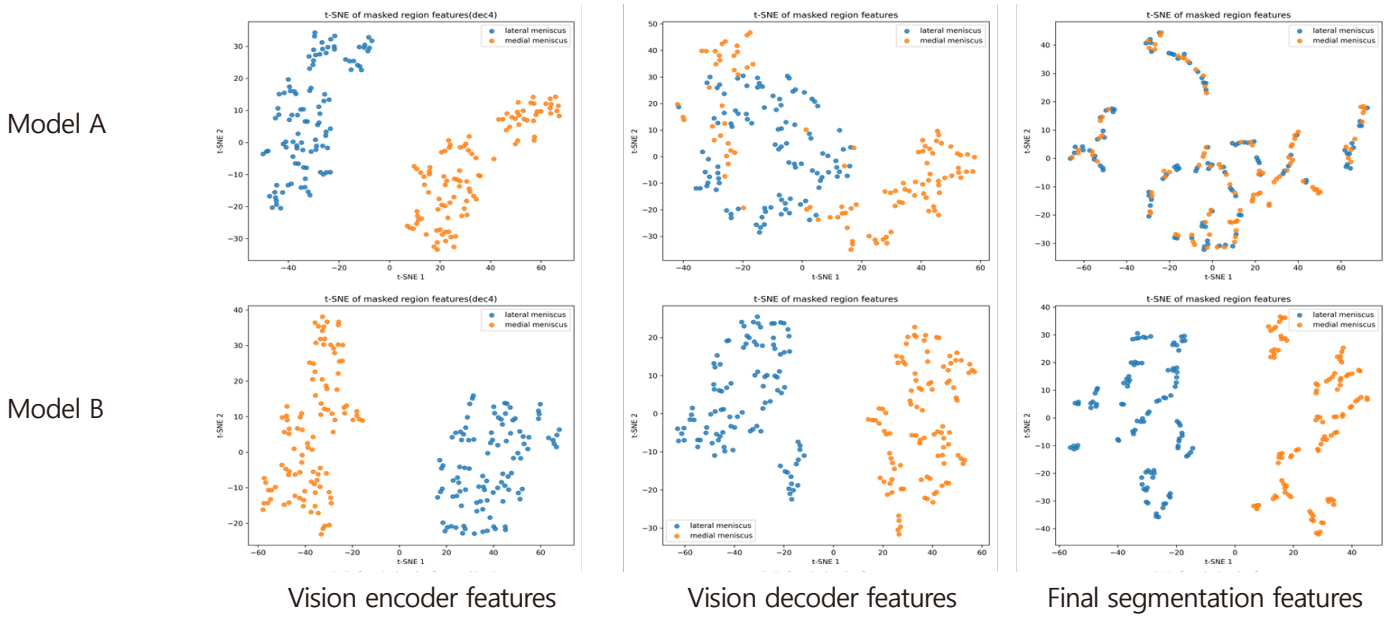

Figure 4. t-SNE visualization of masked region features for lateral and medial meniscus in the latent spaces of the vision encoder, vision decoder and final segmentation layer under different text embedding with LoRA adaptation setting.

Table 2. Quantitative evaluation results of meniscus segmentation according to the number of training samples for each model

(%)

Methods	Total Params	Trainable Params	Training samples			
			10	30	50	83
Model A	62,702,899	107,584 (0.17%)	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	5.76 ± 5.77
Model B	62,702,899	107,584 (0.17%)	28.96 ± 8.20	48.47 ± 7.06	51.18 ± 6.35	73.55±4.98
Model C	62,690,453	19,597,922 (31.26%)	74.09 ± 5.18	83.67 ± 4.01	88.27 ± 2.02	89.89 ± 2.03
Model D	62,780,412	19,782,921 (31.51%)	**86.20 ± 2.37**	**90.21 ± 2.50**	**91.11 ± 2.93**	**91.17 ± 2.37**

보인다.

Text Branch의 LoRA가 특징 표현에 미치는 영향을 분석하기 위해, 그림 4에서는 CLIP 텍스트 인코더 끝단의 MLP에 대한 LoRA 적용 전후의 잠재 표현 공간을 t-SNE를 이용하여 시각화하였다. 분석 결과, 이미지 인코더 단계에서는 LoRA 적용 전후에 특징 분포의 차이가 크지 않은 것으로 나타났다. 반면 이미지 모듈의 디코더 단계에서는 뚜렷한 차이가 관찰되었다. Model A와 Model B는 동일한 디코더 구조를 사용함에도 불구하고 서로 다른 잠재 공간 분포를 보였다. Model A의 경우 외측 및 내측 반월상 연골의 특징 분포가 부분적으로 중첩되는 경향이 나타났으며, 이는 두 구조를 구분하기 위한 표현력이 제한적일 수 있음을 시사한다. 반면 Model B에서는 디코더 단계에서 두 구조의 특징 분포가 보다 명확하게 분리되는 것을 확인할 수 있었다. 또한 최종 분할 특징 공간에서도 이러한 분리 경향이 유지되었다. 이는 CLIP 텍스트 인코더의 MLP에 대한 LoRA 적응이 반월상 연골의 의미적 표현을 보다 효과적으로 형성하고, 해부학적으로 유사한 구조를 구분하기 위한 판별적 특징 공간(discriminative feature space)을 형성하도록 유도했기 때문으로 판단된다. 다만 본 t-SNE 분석은 실제 예측 결과가 아닌 ground-truth mask 영역 기반 특징을 사용하여 수행되었으므로, 실제 분할 결과와 완전히 동일하게 해석하기에는 한계가 있다. 그럼에도 불구하고 본 분석은 text branch의 LoRA가 latent representation 형성에 미치는 영향을 정성적으로 확인할 수 있다는 점에서 의미가 있다.

표 2는 학습 데이터 수에 따른 분할 성능 변화를 분석한 결과를 나타낸다. 전반적으로 모든 모델에서 학습 데이터의 수가 증가함에 따라 분할 성능이 점진적으로 향상되는 경향을 보였다. 특히 제안 방법인 Model D는 데이터가 제한적인 환경에서도 우수한 성능을 나타냈다. 단 10개의 학습 데이터만을 사용한 경우에도 평균 85.21%의 DSC를 달성하였으며, 이는 동일 조건의 다른 모델들보다 현저히 높은 성능이다. 이러한 결과는 대규모 데이터로 사전 학습된 멀티모달 파운데이션 모델을 LoRA를 통해 효율적으로 적응시킴으로써, 소량의 학습 데이터만으로도 반월상 연골의 핵심적인 해부학적 특징을 효과적으로 학습할 수 있음을 보여준다. 또한 Model D는 30개의 학습 데이터만으로도 90.21%의 DSC를 달성하였으며, 이후 학습 데이터 수가 증가함에 따라 성능이 점진적으로 향상되어 83개의 학습 데이터에서 91.17%의 DSC를 기록하였다. 이는 제안 방법이 제한된 데이터 환경에서도 안정적인 성능을 유지할 뿐만 아니라, 학습 데이터가 증가할 경우 추가적인 성능 향상 또한 가능함을 시사한다.

4. 결론

본 논문에서는 대규모 복부 CT 데이터로 사전 학습된 멀티모달 파운데이션 모델과 매개변수 효율적 미세조정 기법인 LoRA를 결합하여, 데이터가 제한적인 환경에서도 무릎 MR 영상 내 반월상 연골을 안정적으로 분할할 수 있는 방법을 제안하였다. 제안된 방법은 영상 정보뿐만 아니라 해부학적 의미를 담은 텍스트 프롬프트를 활용하고, 이미지 및 텍스트 모듈에 LoRA를 적용함으로써 제한된 데이터 환경에서도 효율적인 멀티모달 적응이 가능함을 보였다.

실험 결과, 제안 모델은 전체 파라미터의 31.51%만을 학습에 활용하면서도 외측 및 내측 반월상 연골 모두에서 90% 이상의 높은 DSC를 달성하였다. 또한 추가적인 비라벨 데이터 없이도 기존 반지도학습 기반 방법과 유사한 수준의 성능을

보임으로써, 파운데이션 모델 기반 적응의 효과를 확인하였다. 특히 단 10개의 극소량 학습 데이터만으로도 외측과 내측 반월상 연골에서 각각 84.87%와 85.55%의 다이스 계수를 기록함으로써, 라벨링 비용이 높은 의료 영상 환경에서도 제안 방법이 실용적으로 활용될 수 있음을 보여준다. 이러한 결과는 대규모 CT 데이터로 사전 학습된 멀티모달 파운데이션 모델이 LoRA 기반 적응을 통해 데이터가 제한적인 MR 영상 분할 문제에도 효과적으로 활용될 수 있음을 시사한다.

향후 연구에서는 반월상 연골과 주변 해부학적 구조를 포함한 다중 구조 분할로 연구 범위를 확장하고, 형태학적 특성과 영상 기반의 정보를 반영한 텍스트 프롬프트를 추가적으로 활용함으로써, 실제 임상 환경에서의 실용성과 범용성을 극대화할 수 있는 방향으로 연구를 확장하고자 한다.

감사의 글

이 성과는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원(RS-2022-KH129703) 및 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2024-00340610) 및 서울여자대학교 학술연구비의 지원(2026)을 받아 수행되었으며 이에 감사드립니다.

References

[1] D.J. Hunter, Y.Q. Zhang, J.B. Niu, X. Tu, S. Amin, M. Clancy, and D.T. Felson, "The association of meniscal pathologic changes with cartilage loss in symptomatic knee osteoarthritis." Arthritis & Rheumatism, Vol.54, No.3, pp.795-801, 2006.

[2] J.J. Elsner, S. Portnoy, F. Guilak, A. Shterling, and E. Linder-Ganz, “MRI-based characterization of bone anatomy in the human knee for size matching of a medial meniscal implant.” Journal of biomechanical engineering vol. 132,10, 2010

[3] H. Jeong, J.H. Wang, H. Hong, "Semi-supervised meniscus segmentation with adaptive weighting strategies in a limited labeled data environment." Journal of the Korea Computer Graphics Society, Vol. 31, No. 2, pp. 27–36, 2025.

[4] Ronneberger O, Fischer P, Brox T. "U-Net: Convolutional networks for biomedical image segmentation." Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, pp. 234–241, 2015.

[5] Norman B, Pedoia V, Majumdar S. "Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry." Radiology, Vol. 288, No. 1, pp. 177–185, 2018.

[6] U. Jeon, H. Kim, H. Hong, and J. Wang, "Automatic Meniscus Segmentation Using Adversarial Learning-Based Segmentation Network with Object-Aware Map in Knee MR Images," Diagnostics, Vol. 11, No. 9, pp. 1612, 2021.

[7] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[8] Mazurowski MA, et al. "Segment Anything Model for medical image analysis: an experimental study." Medical Image Analysis, Vol. 89, pp. 102918, 2023.

[9] Liu J, et al. "CLIP-driven universal model for organ segmentation and tumor detection." Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[10] Hu EJ, Shen Y, Wallis P, et al. "LoRA: Low-rank adaptation of large language models." International Conference on Learning Representations (ICLR), 2022.

[11] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[12] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F.Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K.H. Maier-Hein, "nnU-Net: Self-adapting framework for u-net-based medical image segmentation." arXiv preprint arXiv:1809.10486, 2018.

〈저자소개〉

조은빈

- 2026년 8월 서울여자대학교 소프트웨어융합학과 졸업(학사)
- 관심분야 : 의료 인공지능, 딥러닝, 컴퓨터 비전
- https://orcid.org/0009-0007-2224-4021

이민진

- 2007년 2월 서울여자대학교 컴퓨터학과 졸업(학사)
- 2016년 8월 서울여자대학교 컴퓨터학과 졸업(박사)
- 2016년 9월~현재 서울여자대학교 소프트웨어학과 초빙강의교수
- 관심분야 : 의료 인공지능, 딥러닝, 영상 분할 및 분석
- https://orcid.org/0000-0002-6773-1364

왕준호

- 1994년 고려대학교 의학과 학사
- 2002년 고려대학교 정형외과학 석사
- 2004년 고려대학교 정형외과학 박사
- 2002년-2003년 삼성서울병원 전임의
- 2008년-2009년 미국 피츠버그 대학교연구원
- 2004년-2010년 고려대학교 안산병원 조/부교수
- 2010년-현재 삼성서울병원 교수
- 관심분야: 반월상 연골판, 생체 적합 소재, 3D 프린팅
- https://orcid.org/0000-0001-6530-795X

홍헬렌

- 1994년 2월 이화여자대학교 전자계산학과 졸업(학사)
- 1996년 2월 이화여자대학교 전자계산학과 졸업(석사)
- 2001년 8월 이화여자대학교 컴퓨터학과 졸업(박사)
- 2001년 9월~2003년 7월 서울대학교 컴퓨터공학부 BK 조교수
- 2006년 3월~현재 서울여자대학교 소프트웨어학과 교수
- 관심분야 : 의료 인공지능, 딥러닝, 영상처리 및 분석
- https://orcid.org/0000-0001-5044-7909