한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 4, P. 91~100

Support 영상-마스크 쌍 기반 시각 조건 정보를 활용한 복부 CT 영상에서 신장, 종양, 낭종 범용 분할

김고운[O] 이민진[O] 홍헬렌*

서울여자대학교 소프트웨어학과

(gukim, minjin, hlhong)@swu.ac.kr

Universal Segmentation of Kidney, Tumor, and Cyst Using Support Image-Mask Pair-Based Visual Conditioning

Goun Kim[O] Min Jin Lee[O] Helen Hong*

Department of Computer Science, Seoul Women's University

요 약

신장암의 진단 및 수술 계획 수립을 위해서는 복부 CT 영상에서 신장, 신장 종양 및 신장 낭종의 정확한 분할이 중요하다. 기존 CLIP 기반 범용 의료영상 분할 모델은 텍스트 기반 조건 정보를 활용하지만 형태적 다양성과 영상적 이질성이 큰 신장 종양 및 신장 낭종의 시각적 특성을 충분히 반영하기 어렵다. 본 연구에서는 support 영상-마스크 쌍으로부터 클래스 특화 시각 특징을 추출하는 Mask-Guided Sub-pixel Feature Projection(MGSP) 모듈과 종양 크기 기반 support 영상 선택 기법을 결합한 시각 조건 기반 분할 프레임워크를 제안한다. KiTS23 데이터셋 실험 결과 제안 방법은 신장 종양 분할 성능을 DSC 67.46%에서 72.28%로, 신장 낭종 분할 성능을 DSC 53.52%에서 55.13%로 향상시켰으며, 신장 종양 분할의 표준편차를 감소시켜 보다 안정적인 성능을 보였다. 본 연구는 텍스트 기반 조건 정보에 시각 조건 정보를 결합함으로써 형태적 변이가 큰 구조물의 분할 성능을 효과적으로 향상시킬 수 있음을 확인하였다.

Abstract

Accurate segmentation of kidneys, kidney tumors, and renal cysts in abdominal CT images is essential for the diagnosis and surgical planning of renal cell carcinoma. Existing CLIP-based universal medical image segmentation models use text-based conditional information, however, they have limitations in capturing the visual characteristics of tumors and cysts with high morphological diversity and imaging heterogeneity. In this study, we propose a visual-conditioned segmentation framework that integrates a Mask-Guided Sub-pixel Feature Projection (MGSP) module, which extracts class-specific visual features from support image-mask pairs, and a tumor size-aware support selection strategy. Experimental results on the KiTS23 dataset demonstrate that the proposed method improves kidney tumor segmentation performance from 67.46% to 72.28% in DSC and cyst segmentation performance from 53.52% to 55.13% in DSC, while reducing the standard deviation of tumor segmentation, indicating more stable performance. These findings indicate that integrating visual condition information with text-based conditional information can effectively improve segmentation performance for structures with high morphological variability.

키워드: 신장 및 병변 분할, 범용 의료영상 분할, 클래스 특화 특징 벡터, support 영상-마스크 쌍, 종양 크기 기반 support 영상 선택

Keywords: Kidney and lesion segmentation, Universal medical image segmentation, Class-specific feature vector, Support image-mask pair, Tumor size-aware support selection

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

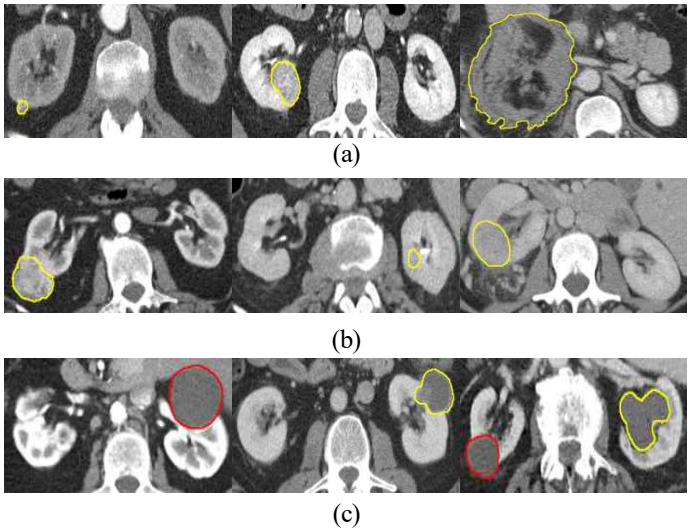
*corresponding author: Helen Hong / Department of Computer Science, Seoul Women's University (hlhong@swu.ac.kr)

Received : 2026.06.10./ Review completed : 1st 2026.07.24./ Accepted : 2026.08.09.
DOI : 10.15701/kcgs.2026.32.4.91
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

신장암(Renal Cell Carcinoma)은 전 세계적으로 발생 빈도가 지속적으로 증가하고 있는 대표적인 비뇨기계 악성 종양으로, 진단과 수술 계획 수립을 위해 복부 전산화단층촬영(Computed Tomography, CT)이 널리 활용되고 있다[1,2]. 최근에는 정상 신장 기능을 최대한 보존하면서 신장 종양만을 제거하는 부분신장절제술(Partial Nephrectomy)이 표준 치료로 자리잡고 있으며[3], 이를 안전하게 수행하기 위해서는 신장 종양의 크기, 신장 내 위치 및 집합계(Collecting System)와의 해부학적 관계를 정량적으로 평가하는 과정이 필수적이다. 임상에서는 이러한 수술 난이도 평가를 위해 RENAL nephrometry score 가 널리 활용되고 있으며[4], 이를 자동 산출하기 위해서는 CT 영상에서 신장(Kidney), 신장 종양(Tumor), 신장 낭종(Cyst)에 대한 정확한 분할이 선행되어야 한다. 특히 신장 낭종은 CT 영상에서 신장 종양과 유사한 영상 밝기값 및 형태를 보이는 경우가 많아 오진 및 오분할의 원인이 될 수 있으므로 신장 종양과 구분하여 분석하는 것이 중요하다[5].

그러나 그림 1 과 같이 신장 종양은 환자마다 크기와 형태가 다양하며, 내부 괴사(necrosis), 출혈(hemorrhage) 및 낭성 변화(cystic degeneration)등으로 인해 영상적 이질성(heterogeneity)이 크게 나타난다. 또한 신장 낭종은 크기와 형태가 다양하고 일부 경우 신장 종양과 유사한 영상 특성을 보이기 때문에 두 병변을 명확하게 구분하는 것은 쉽지 않다. 이러한 특성으로 인해 신장, 종양 및 신장 낭종을 동시에 정확하게 분할하는 것은 여전히 어려운 문제이며, 복잡한 시각적 변이를 효과적으로 반영할 수 있는 분할 방법이 요구된다.



(a) (b) (c)

Figure 1. Challenges in kidney tumor segmentation: (a) Diverse tumor locations, sizes, shapes, and heterogeneous appearances, (b) Similar intensity distributions between kidney tumors and surrounding renal parenchyma, and (c) Similar imaging characteristics between kidney tumors and renal cysts. The yellow and red contours represent the kidney tumor and cyst, respectively.

의료 영상 분할 연구는 주로 지도학습 기반 분할(supervised segmentation)을 중심으로 발전해왔다. 초기에는 인코더-디코더 구조와 스킵 연결(skip connection)을 이용한 U-Net[6] 계열의 합성곱 신경망(Convolutional Neural Network, CNN)이 의료영상 분할에서 우수한 성능을 보였다[7,8]. 그러나 CNN 기반 방법은 제한된 수용 영역(receptive field)으로 인해 국소 특징(local feature) 추출에는 효과적이지만 영상 전체의 장거리 의존성(long-range dependency)을 반영하는데에는 한계가 있다. 이를 해결하기 위해 트랜스포머(Transformer) 기반 구조가 도입되었으며, self-attention 메커니즘을 활용하여 전역 문맥(global context)와 장거리 공간 관계를 효과적으로 반영함으로써 복잡한 해부학적 구조의 분할 성능을 향상시키고자 하였다[9-11]. 최근에는 특정 장기나 질환에 특화된 분할 모델을 넘어 다양한 장기와 병변을 하나의 모델에서 통합적으로 처리하는 범용 의료영상 분할(Universal Medical Image Segmentation) 연구가 활발히 진행되고 있다[12-15]. 이러한 방법들은 장기 또는 병변의 명칭을 텍스트 프롬프트로 입력하거나 학습 가능한 프롬프트(prompt token)을 활용하여 클래스 정보를 조건(condition)으로 제공한다. 그러나 텍스트 기반 조건 정보는 “신장”, “신장 종양”, “신장 낭종”과 같은 클래스 수준의 의미 정보는 제공할 수 있으나 동일한 클래스 내에서도 환자별로 크게 달라지는 형태, 크기 및 내부 조직 특성과 같은 시각적 외형 정보를 충분히 반영하지 못한다는 한계가 있다. 특히 신장 종양과 신장 낭종은 동일 클래스 내 변이가 크고 유사한 영상 패턴을 나타내므로 클래스 명칭만으로는 두 병변을 안정적으로 구분하기 어렵다.

본 연구에서는 텍스트 기반 조건 정보만으로는 충분히 반영하기 어려운 신장, 신장 종양 및 신장 낭종의 형태적, 영상학적 특성을 분할 과정에 직접 활용하기 위해 support 영상-마스크 쌍(image-mask pair) 기반의 시각 조건 정보(visual condition)를 이용하는 새로운 조건부 분할 프레임워크를 제안한다. 제안 방법은 support 영상으로부터 추출한 클래스 특화 시각 특징(class-specific visual representation)을 분할 네트워크의 조건 정보로 활용함으로써, 텍스트 프롬프트가 제공하는 클래스 수준의 의미 정보와 실제 구조물의 시각적 외형 정보를 동시에 반영한다. 이를 위해 support 마스크를 이용하여 관심 구조물의 전경 특징을 선택적으로 추출하는 Mask-Guided Sub-pixel Feature Projection(MGSP) 모듈을 설계하고 생성된 클래스 특화 특징 벡터를 분할 과정에 효과적으로 통합한다. 또한, query 영상과 유사한 신장 종양 크기 특성을 갖는 support 영상-마스크 쌍을 선택하는 종양 크기 기반 참조 영상 선택 기법(Tumor Size-aware Support Selection Strategy)을 제안하여 support-query 간 시각적 불일치를 최소화하고 보다 적절한 시각적 사전정보(visual prior)를 제공한

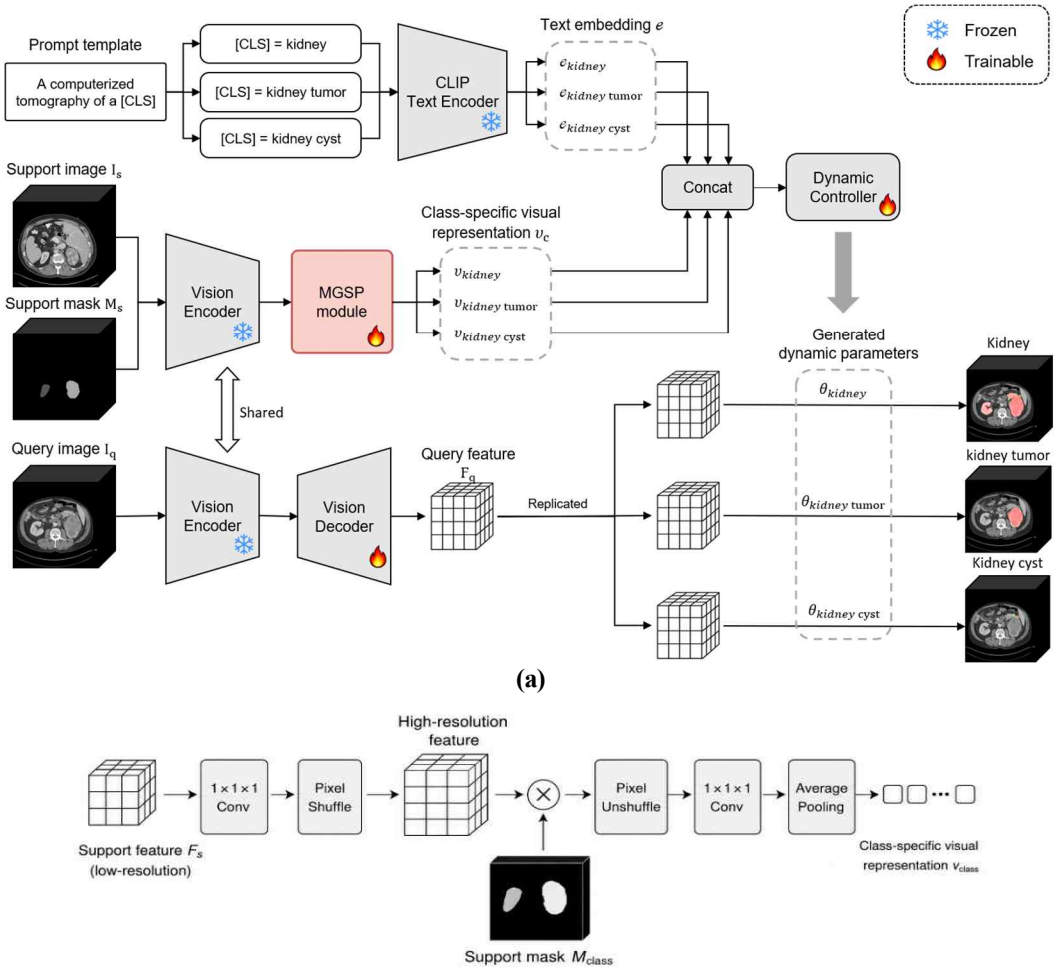


Figure 2. Overview of the proposed visual-conditioned kidney structure segmentation framework (a) Overall architecture of the proposed framework, which generates class-specific dynamic segmentation parameters by combining CLIP text embeddings with visual representations extracted from support image-mask pairs, and (b) Generation process of class-specific visual representations using the proposed Mask-Guided Sub-pixel Feature Projection (MGSP) module.

다. 이를 통해 신장, 신장 종양 및 신장 낭종을 동시에 분할하는 다중 구조물 분할 문제에서 시각 조건 정보의 효과를 검증하고, 기존 텍스트 기반 범용 의료영상 분할 기법의 한계를 보완할 수 있는 새로운 시각 조건 기반 분할 방법을 제안한다.

2. 제안 방법

그림 2(a)는 본 연구에서 제안하는 프레임워크의 전체 구조를 나타낸다. 제안 방법은 query 영상과 support 영상-마스크 쌍을 입력으로 받아 클래스 특화 특징 벡터를 생성하고, 이를 텍스트 임베딩과 결합하여 동적 분할 파라미터를 생성한다. 생성된 파라미터는 query 영상의 특징 맵에 적용되어 신장, 신장 종양 및 신장 낭종의 분할 결과를 생성한다.

2.1 CLIP 기반 범용 의료영상 분할 모델

본 연구는 CLIP-Driven Universal Model을 기반 분할 네트워크로 사용한다. 해당 모델은 영상 특징을 추출하는 이미지 모듈, 클래스 의미 정보를 생성하는 텍스트 모듈, 그리고 두 정보를 결합하여 클래스별 분할 파라미터를 생성하는 동적 제어기(dynamic controller)로 구성된다.

이미지 모듈은 SwinUNETR를 기반으로 구성되며, 입력 CT 영상으로부터 다중 스케일 특징을 추출한 후 디코더를 통해 입력 영상과 동일한 해상도의 특징맵을 생성한다. 텍스트 모듈은 CLIP 텍스트 인코더를 이용하여 "A computerized tomography of a {CLS}" 형태의 프롬프트를 임베딩 벡터로 변환한다. 여기서 {CLS}는 신장(Kidney), 신장 종양(Tumor), 신장 낭종(Cyst)과 같은 분할 대상 클래스명을 의미한다.

동적 제어기는 이미지 특징과 텍스트 임베딩을 결합하여 클래스별 동적 파라미터 θ를 생성한다. 생성된 파라미터는 최종 특징 맵에 적용되어 각 클래스에 대한 분할 마스크를 생성한다. 그러나 기존 방법은 텍스트 임베딩만을 조건 정보로 사용하기 때문에 동일 클래스 내에서 나타나는 다양한 형태적 및 영상학적 특성을 충분히 반영하기 어렵다. 특히 신장 종양과 신장 낭종은 환자 간 형태 변이가 크고 일부 경우 유사한 영상 패턴을 나타내므로 추가적인 시각 조건 정보가 필요하다.

2.2 MGSP 기반 클래스 특화 특징 벡터 생성

본 연구에서는 텍스트 기반 조건 정보의 한계를 보완하기 위해 support 영상-마스크 쌍으로부터 클래스 특화 시각 특징(class-specific visual representation)을 추출하는 MGSP(Mask-Guided Sub-pixel Feature Projection) 모듈을 제안한다.

Support 영상은 이미지 인코더를 통해 특징 맵 F_s로 변환된다. 일반적으로 인코더 출력 특징은 입력 영상보다 낮은 공간 해상도를 가지므로 support 마스크를 직접 적용할 경우 작은 종양이나 복잡한 경계 구조의 공간 정보가 손실될 수 있다. 이를 해결하기 위해 MGSP는 픽셀 셔플(Pixel Shuffle)과 픽셀 역셔플(Pixel Unshuffle)을 활용하여 마스크의 공간 정보를 유지하면서 전경 특징을 선택적으로 추출한다.

그림 2(b)와 같이 먼저 support 특징 F_s에 1x1x1 컨볼루션을 적용한 후 픽셀 셔플 연산을 통해 support 마스크와 동일한 해상도로 확장한다. 이후 클래스별 마스크 M_{class} 를 적용하여 관심 구조물에 해당하는 특징만을 선택적으로 유지한다. 마스크가 적용된 특징은 픽셀 역셔플 연산을 통해 원래 해상도로 복원되며, 추가적인 1x1x1 컨볼루션을 거쳐 전역 평균 풀링(Global Average Pooling)을 수행함으로써 클래스 특화 특징 벡터를 생성한다.

$$v_{class} = GAP\left(Conv\left(PU\left(PS\left(Conv\left(F_s\right)\right) \odot M_{class}\right)\right)\right) \quad (1)$$

여기서 F_s는 support 영상 맵을 의미하며, PS(·)와 PU(·)는 각각 픽셀 셔플과 픽셀 역셔플 연산을 나타낸다. 또한 M_{class} 는 클래스별 support 마스크를 의미한다.

생성된 클래스 특화 특징 벡터 v_{class} 은 CLIP 텍스트 임베딩 t_{class} 와 결합되어 식(2)와 같이 동적 제어기의 조건 정보로 사용된다.

$$p_{class} = Conv\left([v_{class} \,||\, t_{class}]\right) \quad (2)$$

여기서 $[\cdot \,||\, \cdot]$ 는 채널 방향 결합(concatenation)을 의미하며, $Conv$ 는 1x1x1 컨볼루션으로 구성된 동적 제어기를 의미한다. 이를 통해 클래스의 의미 정보뿐만 아니라 실제 구조물의 형태적 및 영상학적 특성이 동적 파라미터 생성 과정에 반영될 수 있다.

2.3 종양 크기 기반 Support 영상 선택

Support 영상으로부터 생성되는 시각 특징은 선택된 support 영상의 특성에 크게 영향을 받는다. 특히 신장 종양은 환자마다 크기와 형태의 변동성이 매우 크므로 query 영상과 support 영상 간의 시각적 불일치가 발생할 수 있다. 이를 완화하기 위해 본 연구에서는 신장 종양 부피를 기준으로 데이터를 소형(small), 중형(medium), 대형(large), 극대형(extremely large) 네 개의 크기 그룹으로 구분하고[16], 학습 과정에서는 query 영상과 동일한 크기 그룹에 속하는 support 영상-마스크 쌍을 무작위로 선택하여 사용한다. 이를 통해 query와 support 간의 시각적 유사성을 높이고 보다 적절한 시각 조건 정보를 제공할 수 있다.

추론 단계에서는 query 영상의 신장 종양 크기를 사전에 알 수 없으므로 각 크기 그룹의 대표 support 영상-마스크 쌍을 이용하여 독립적인 분할 결과를 생성한다. 이후 각 복셀에서 가장 높은 두 개의 예측 확률을 선택하여 평균하는 Top-2 Probability Aggregation 기법을 적용함으로써 최종 분할 결과를 생성한다. 이를 통해 특정 support 조건에 대한 의존성을 감소시키고 보다 안정적이고 강건한 분할 결과를 얻을 수 있다.

3. 실험 및 결과

3.1 실험 데이터 및 환경

본 연구에서는 2023 Kidney and kidney Tumor Segmentation (KiTS23) Challenge에서 제공된 공개 데이터셋[17-19]를 사용하였다. 해당 데이터셋은 2010년부터 2018년까지 미네소타 대학 의료 센터에서 신장 절제술을 받은 환자 중 무작위로 선택된 589명의 복부 CT 영상으로 구성되어 있다. 제공된 CT 영상은 512x512 픽셀 해상도를 가지며, 평면 내 해상도(in-plane resolution)는 0.5mm~5.0mm, 슬라이스 두께는 0.4375mm~1.041mm범위로 구성되어 있다. 기반 모델의 사전학습에 사용된 KiTS19 데이터셋과 중복되지 않는 KiTS23 데이터만을 사용하였으며, 신장 종양과 신장 낭종 마스크가 중복 표기되거나 일부 신장 낭종 마스크가 누락된5개의 데이터를 제외한 총 184개의 데이터를 사용하였다. 이 중 137개는 훈련 데이터셋으로, 47개는 테스트 데이터셋으로 사용하였다. 또한 support 영상-마스크 쌍은 신장, 신장 종양 및 신장 낭종의 클래스 특화 특징 정보를 생성하기 위해 신장 낭종이 포함된 훈련 데이터에서 구성하였다. 이는 모든 클래스에 대한 시각 조건 정보를 생성하기 위함이다.

본 연구에서는 신장, 신장 종양, 신장 낭종 및 주변 연부조직의 영상 정보를 효과적으로 반영하기 위해 CT 밝기값을 -175 HU에서 250 HU 범위로 제한(clipping)한 후, 최소-최대 정규화(min-max normalization)를 적용하여 0~1범위로 변환하였다. 또한 영상 해상도의 일관성을 확보하기 위해 모든 영상을 1.5mm등방성 해상도로 재표본화하였다.

제안된 네트워크를 학습하고 평가하기 위해 NVIDIA GeForce RTX 3090 GPU, Intel Xeon CPU, 32GB RAM이 장착된 Ubuntu 20.04.6 LTS 운영체제의 서버에서 수행하였고, CUDA 12.6 버전의 GPU 환경에서 파이썬(python) 3.7 및 파이토치(pytorch) 1.11.0 라이브러리를 사용해 진행하였다. 제안 방법은CLIP-Driven

Universal Model의 공개 코드를 기반으로 구현되었다. 본 연구에서는 모델 평가를 위해 홀드아웃 (hold-out) 방식을 사용하였다. 하이퍼파라미터로는 배치(batch) 크기 1로 설정하였고, 최적화에는 AdamW 옵티마이저를 사용하였다. 사전학습된 영상 인코더와 텍스트 인코더의 파라미터는 고정하였으며, 사전 학습된 모듈인 영상 디코더와 동적 제어기는 낮은 학습률인 1×10^{-5} 를 적용하였으며, 새롭게 추가된 MGSP 모듈은 충분한 학습을 위해 상대적으로 높은 1×10^{-3}의 학습률을 적용하였다. 또한 학습 초기의 불안정성을 완화하기 위해 linear warm-up 을 적용한 뒤, cosine annealing방식으로 학습률을 점진적으로 감소시키는 학습률 스케쥴러(learning rate scheduler)를 사용하였다. 학습은 최대 500에폭으로 설정하였으나, 학습 손실이 수렴하는 약 300 에폭에서 조기 종료하였다. 손실 함수로는 다이스 손실함수(Dice loss)와 이진 교차 엔트로피 손실 함수를 사용하여 분할 성능을 최적화하였다.

제안 방법의 성능은 정량적 및 정성적 측면에서 평가하였다. 정량적 평가는 다이스 유사 계수(Dice Similarity Coefficient, DSC), 정밀도(Precision), 재현율(Recall), 균형 정확도(Balanced accuracy, Bal. Acc)를 기준으로 평가하였다. 신장, 신장 종양, 신장 낭종의 종합적인 분할 성능을 효과적으로 평가하기 위해, 비교 방법으로 영상 정보만을 사용하는 SwinUNETR과 영상-텍스트 정보를 활용하는 CLIP-Driven Universal Model을 선정하였다. SwinUNETR은 CLIP-Driven Universal Model의 영상 특징 추출 네트워크로 사용되는 모델로, CLIP 기반 조건 정보의 효과를 분석하기 위한 기준 모델로 채택하였다. 또한 제안 방법의 각 구성 요소의 효과를 분석하기 위해 support 영상만 조건 정보로 사용하는 방법(Support image condition), support 영상-마스크 쌍을 사용하는 방법(Support image-mask pair condition), 그리고 종양 크기 기반 support 영상 선택을 추가한 제안 방법(Support image-mask pair condition + Size-aware)을 함께 비교하였다. 종양 크기별 성능 비교에서는 수술 난이도 평가에 핵심적인 요소인 신장 종양만을 대상으로 성능을 추가 분석하였다. 이를 위해 단일 장기 분할에서 우수한 성능을 보이는 대표적인 최신 분할 모델인 nnU-Net을 추가적인 비교 대상으로 선정하였다. 비교 방법들인 SwinUNETR과 nnU-Net은 모든 파라미터를 학습하는 완전 미세조정 방식으로 학습하였으며, CLIP-Driven Universal Model과 제안 방법은 공개된 사전학습 가중치를 초기화 값으로 사용하여 학습을 수행하였다.

Table 1. Quantitative comparison of segmentation performance for kidney, tumor, and cyst structures. Values are presented as mean ± standard deviation, and the best performance for each metric is shown in bold.

Structure	Method	DSC (%)	Precision (%)	Recall (%)	Bal. Acc. (%)
Kidney	SwinUNETR	82.70±23.42	95.46±13.31	77.14±25.53	88.55±12.78
	CLIP-Driven-Universal Model (baseline)	95.23±3.97	96.14±3.57	94.52±5.66	97.24±2.84
	Support image condition	95.31±3.42	**96.30±2.68**	94.54±5.48	97.25±2.75
	Support image-mask pair condition	95.35±3.42	95.88±3.97	95.00±4.47	97.48±2.25
	Support image-mask pair condition + Size-aware	**95.53±2.90**	95.92±3.15	**95.28±4.23**	**97.62±2.13**
Tumor	SwinUNETR	41.20±32.70	56.69±35.82	38.82±33.32	69.39±16.66
	CLIP-Driven-Universal Model (baseline)	67.46±26.18	74.89±27.26	67.41±27.65	83.68±13.82
	Support image condition	71.18±22.00	74.05±24.40	73.08±23.67	86.52±11.83
	Support image-mask pair condition	71.77±22.56	73.58±25.92	**74.96±23.53**	**87.45±11.76**
	Support image-mask pair condition + Size-aware	**72.28±22.42**	**75.68±25.77**	73.51±23.47	86.73±11.73
Cyst	SwinUNETR	30.62±26.42	64.58±35.93	25.66±26.06	62.83±13.03
	CLIP-Driven-Universal Model (baseline)	53.52±30.95	60.08±32.77	53.45±30.73	76.72±15.36
	Support image condition	46.95±35.98	57.93±38.89	46.39±36.08	73.19±18.04
	Support image-mask pair condition	52.95±34.59	61.94±36.42	52.30±36.49	76.15±18.24
	Support image-mask pair condition + Size-aware	**55.13±34.85**	**62.65±36.20**	**54.26±35.74**	**77.13±17.87**

3.2 정량적 평가 결과

표 1 은 신장, 신장 종양 및 신장 낭종에 대한 정량적 분할 성능을 나타낸다. 전체적으로 CLIP 기반 방법들은 SwinUNETR 보다 우수한 성능을 보였으며, 특히 형태적 변동성과 영상적 이질성이 큰 신장 종양 및 신장 낭종 분할에서 성능 향상이 두드러졌다. 이는 텍스트 기반 의미 정보가 복잡한 병변의 시각적 특성을 학습하는 데 효과적으로 활용될 수 있음을 보여준다. 기반 모델인 CLIP-Driven Universal Model 과 비교할 때, 제안 방법은 support 기반 시각 조건 정보를 추가함으로써 신장 종양 및 신장 낭종 분할 성능을 지속적으로 향상시켰다. 특히 support 영상만을 사용하는 경우보다 support 영상-마스크 쌍을 사용하는 경우 더 높은 성능을 보였으며, 이는 support 마스크가 관심 구조물에 해당하는 전경 특징만을 선택적으로 추출하여 보다 정확한 클래스 특화 특징 표현을 생성할 수 있기 때문으로 해석된다.

신장 분할의 경우 구조적 형태가 비교적 일정하고 경계가 명확하기 때문에 모든 CLIP 기반 방법이 유사한 성능을 보였다. 반면 신장 종양과 신장 낭종은 크기와 형태의 다양성이 크고 주변 조직과의 경계가 불명확하므로 시각 조건 정보의 효과가 더욱 크게 나타났다. 신장 종양 분할에서는 support 영상만을 조건 정보로 활용하였을 때 DSC 가 67.46%에서 71.18%로 향상되었으며, support 영상-마스크 쌍을 적용한 경우 71.77%까지 추가 향상되었다. 이는 MGSP 모듈이 신장 종양 영역에 해당하는 전경 특징을 효과적으로 추출함으로써 종양 관련 시각 정보를 보다 정확하게 반영할 수 있음을 의미한다. 또한 종양 크기 기반 support 선택 기법을 적용한 최종 모델은 DSC 72.28%, Precision 75.68%를 기록하여 가장 우수한 성능을 보였다. 특히 Precision 향상은 query 영상과 유사한 크기의 신장 종양 정보를 활용함으로써 종양과 주변 조직 간의 혼동을 감소시킨 결과로 판단된다. 신장 낭종 분할에서는 support 영상만 사용한 경우 DSC 가 53.52%에서 46.95%로 감소하였다. 이는 support 영상 전체가 신장 종양과 신장 낭종이 혼재된 시각 정보를 포함하고 있어 클래스 특이성이 부족하기 때문으로 해석된다. 반면 support 영상-마스크 쌍을 적용하면 신장 낭종 영역에 해당하는 특징만 선택적으로 활용할 수 있으므로 DSC 가 52.95%로 회복되었으며, 종양 크기 기반 support 선택 기법을 추가한 경우 최종적으로 55.13%를 기록하였다. 이러한 결과는 support 마스크가 신장 종양과 신장 낭종 간의 시각적 혼동을 감소시키는 데 중요한 역할을 수행함을 보여준다. 또한 신장 종양 분할의 경우 DSC 의 표준편차가 26.18 에서 22.42 로 감소하였다. 이는 제안 방법이 평균 성능뿐만 아니라 다양한 신장 종양 크기와 형태에 대해서도 보다 안정적인 분할 결과를

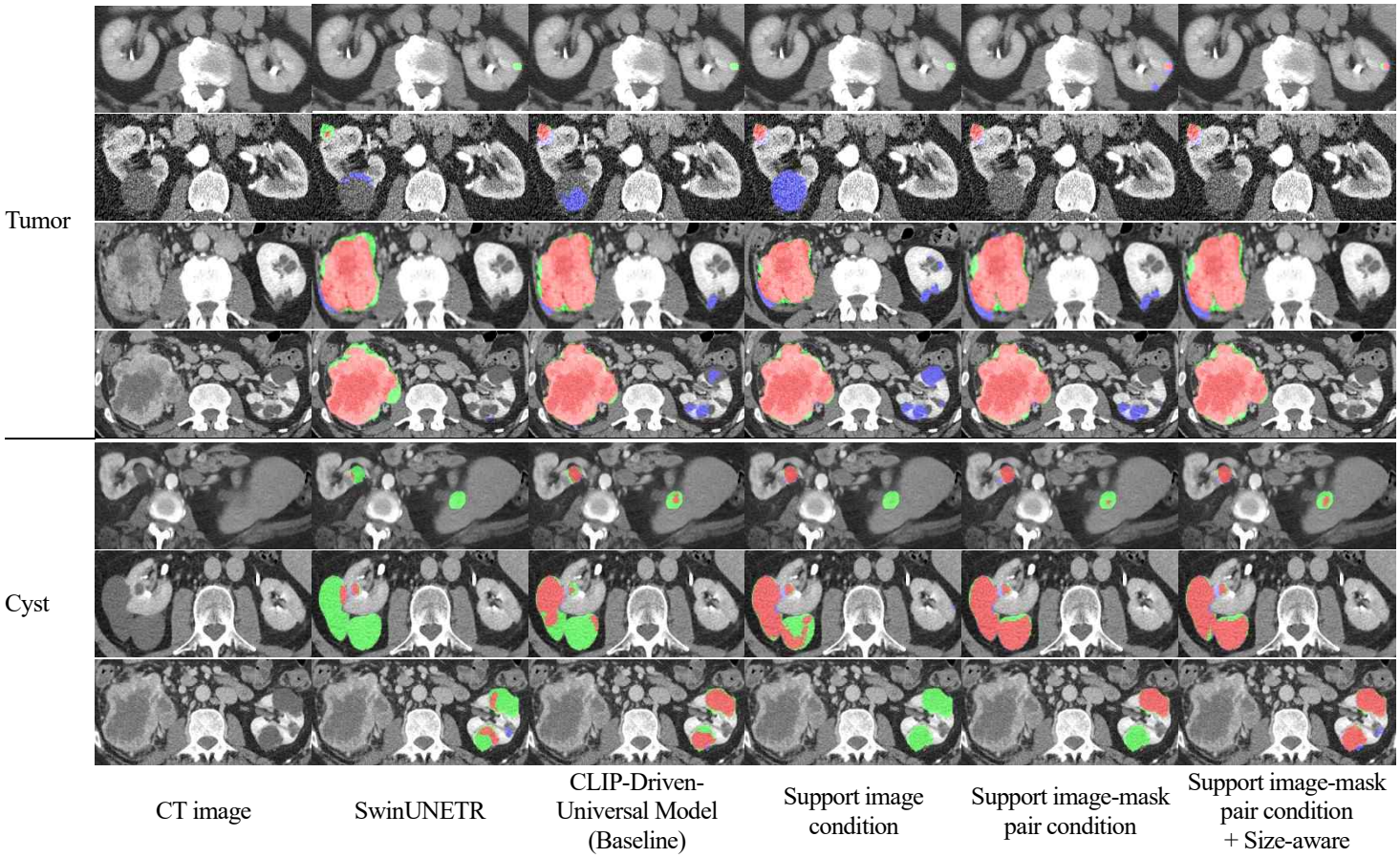


Figure 3. Qualitative evaluation of kidney tumor and cyst segmentation results in abdominal CT images. The red, green, blue colors represent true positive, false negative(under-segmentation), false positive(over-segmentation), respectively.

제공할 수 있음을 의미한다. 통계적 유의성은 복부 CT 영상에서 신장 종양 및 신장 낭종이 배경 영역에 비해 상대적으로 작은 비율을 차지하여 전경과 배경 간 클래스 불균형이 존재한다는 점을 고려하여, 균형 정확도를 대상으로 양측 대응 표본 t-검정(two-sided paired t-test)을 수행하였다. 검증 결과, 최종 제안 방법은 비교방법(SwinUNETR 및 CLIP-Driven Universal Model) 대비 신장과 신장 종양에서 모두 통계적으로 유의하게 높은 성능을 보였으며($p<0.05$), 신장 낭종에서는 SwinUNETR 대비 통계적으로 유의한 성능 향상을 보였으나($p < 0.05$), CLIP-Driven Universal Model 과는 통계적으로 유의한 차이가 확인되지 않았다($p > 0.05$).

3.3 정성적 평가 결과

그림 3 은 각 방법에 대한 신장 종양 및 신장 낭종 분할 결과를 시각적으로 비교한 것이다. 전체적으로 SwinUNETR 과 CLIP-Driven Universal Model 은 병변의 일부 영역을 포함하지 못하는 과소분할(under-segmentation) 또는 주변 조직을 병변으로 예측하는 과대분할(over-segmentation)이 빈번하게 발생하였다. 신장 종양 분할 결과에서는 support 조건 정보를 활용함에 따라 종양 경계가 보다 정확하게 복원되는 것을 확인할 수 있다. 특히 첫 번째 및 두 번째 행에서는 MGSP를 이용한 support 영상-마스크 쌍 조건이 신장 종양의 실제 형태를 보다 정확하게 반영하여 과소분할 영역을 감소시켰다. 반면 일부 사례에서는 신장 낭종 영역이 신장 종양으로 오분할되는 현상이 관찰되었 으며, 이는 신장 종양과 신장 낭종이 유사한 영상 특성을 갖기 때문으로 판단된다. 제안 방법은 종양 크기 기반 support 선택 방식을 통해 이러한 신장 종양-낭종 간 혼동을 완화하였으며 잘못 분할된 신장 낭종 영역을 효과적으로 제거하는 결과를 보였다.

신장 낭종 분할 결과에서도 유사한 경향이 관찰되었다. Support 영상만을 사용하는 경우 신장 낭종의 형태 정보가 충분히 반영되지 못하여 과소분할이 증가하는 경향을 보였으나, support 영상-마스크 쌍을 적용하면 신장 낭종의 경계와 내부 구

Table 2. Quantitative comparison of kidney tumor segmentation performance across different tumor size groups.

Size	Methods	DSC (%)	Precision (%)	Recall (%)	Bal. Acc.(%)
Small (tumor < 2.4cm)	nnU-Net	61.71±27.18	**70.58±32.34**	61.04±29.73	80.52±14.86
	CLIP-Driven-Universal Model (baseline)	56.92±31.60	64.86±34.55	53.97±33.16	76.98±16.58
	Support image-mask pair condition + Size-aware	**65.70±25.84**	69.25±28.65	**65.12±28.19**	**82.56±14.09**
Medium (2.4cm ≤ tumor < 4.3cm)	nnU-Net	**78.01±18.37**	**89.60±15.35**	**75.04±22.48**	**87.52±11.24**
	CLIP-Driven-Universal Model (baseline)	66.51±26.07	70.58±28.50	67.44±27.28	83.71±13.64
	Support image-mask pair condition + Size-aware	72.34±19.01	74.95±22.83	74.51±21.42	87.25±10.71
Large (4.3cm ≤ tumor < 7.4cm)	nnU-Net	**78.24±24.22**	**77.03±24.22**	**86.03±11.42**	**92.98±5.69**
	CLIP-Driven-Universal Model (baseline)	67.63±24.28	84.68±17.01	65.04±27.29	82.51±13.64
	Support image-mask pair condition + Size-aware	70.17±26.33	75.25±30.75	68.89±27.42	84.44±13.70
Extremely large (7.4cm ≤ tumor)	nnU-Net	80.82±24.07	**85.87±25.42**	**85.36±19.88**	**92.60±9.91**
	CLIP-Driven-Universal Model (baseline)	79.45±20.52	82.46±23.56	83.17±16.05	91.51±8.04
	Support image-mask pair condition + Size-aware	**81.81±19.91**	82.46±24.49	84.81±14.42	92.33±7.21
Overall	nnU-Net	**75.20±21.96**	**82.09±24.21**	**76.60±23.26**	**88.27±11.61**
	CLIP-Driven-Universal Model (baseline)	67.46±26.18	74.89±27.26	67.41±27.65	83.68±13.82
	Support image-mask pair condition + Size-aware	72.28±22.42	75.68±25.77	73.51±23.47	86.73±11.73

조가 보다 정확하게 복원되었다. 또한 제안 방법은 신장 종양과 신장 낭종의 경계가 인접한 영역에서도 오분류를 감소시켜 가장 안정적인 분할 결과를 나타냈다.

3.4 종양 크기별 성능 분석

표 2 는 신장 종양 크기에 따른 신장 종양 분할 성능을 분석한 결과를 나타낸다. 전체 평균 성능은 nnU-Net 이 가장 높았으나 제안 방법은 소형 종양과 극대형 종양에서 가장 우수한 성능을 보였다.

소형 종양의 경우 제안 방법은 DSC 65.70%, Recall 65.12%를 기록하여 nnU-Net 대비 각각 3.99%p 및 4.08%p 높은 성능을 보였다. 일반적으로 소형 종양은 병변 크기가 작고 경계 정보가 제한적이기 때문에 누락되기 쉽다. 제안 방법은 support 기반 시각 조건 정보를 활용하여 작은 종양의 형태 정보를 보완함으로써 탐지 및 분할 성능을 향상시킨 것으로 판단된다. 극대형 종양에서도 제안 방법은 DSC 81.81%로 가장 우수한 성능을 기록하였다. 이는 support 영상으로부터 추출된 클래스 특화 특징이 크고 복잡한 신장 종양의 다양한 형태적 변이를 효과적으로 반영할 수 있음을 나타낸다.

반면 중형 및 대형 종양에서는 nnU-Net 이 더 높은 성능을 보였다. 이는 해당 크기 범주의 신장 종양이 데이터셋 내에서 상대적으로 높은 빈도로 존재하여 완전 지도학습 기반 모델이 충분한 학습 데이터를 통해 안정적인 형태 표현을 학습할 수 있었기 때문으로 판단된다. 그럼에도 불구하고 제안 방법은 데이터 불균형과 형태적 변이가 큰 소형 및 극대형 종양에서 상대적으로 높은 성능을 보였으며, 이는 support 기반 시각 조건 정보가 일반적인 형태 분포에서 벗어난 어려운 사례(hard cases)의 분할 성능 향상에 기여할 가능성을 보여준다. 다만, 종양 크기별 통계적 유의성 검증 결과에서는 모든 크기 그룹에서 모델 간 성능 차이가 통계적으로 유의하지 않았으며($p>0.05$), 이러한 결과는 특정 방법의 우수성을 의미하기보다는, support 기반 시각 조건 정보가 소형 및 극대형 종양에서 성능 향상에 기여할 가능성을 시사하는 결과로 해석할 수 있다.

4. 결론

본 연구에서는 복부 CT 영상에서 신장, 신장 종양 및 신장 낭종을 동시에 분할하기 위한 시각 조건 기반(visual-conditioned) 의료영상 분할 프레임워크를 제안하였다. 기존 CLIP 기반 범용 의료영상 분할 모델은 텍스트 임베딩을 이용하여 구조물의 의미 정보를 제공하지만 신장 종양 및 신장 낭종과 같이 형태적 다양성과 영상적 이질성이 큰 구조물의 시각적 특성을 충분히 반영하기 어렵다는 한계가 있었다. 이를 해결하기 위해 본 연구에서는 support 영상-마스크 쌍으로부터 관심 구조물의 전경 특징을 효과적으로 추출하는 Mask-Guided Sub-pixel Feature Projection(MGSP) 모듈을 설계하고, 이를 통해 생성된 클래스 특화 특징 벡터를 텍스트 임베딩과 결합하여 동적 분할 파라미터 생성 과정에 활용하였다. 또한 query 영상과 support 영상 간의 시각적 불일치를 완화하기 위해 종양 크기 기반 support 영상 선택 기법을 제안하였다.

KiTS23 데이터셋을 이용한 실험 결과 제안 방법은 신장 분할에서는 기존 CLIP 기반 모델과 유사한 수준의 높은 성능을 유지하면서도, 신장 종양 및 신장 낭종 분할에서는 일관된 성능 향상을 보였다. 특히 support 영상-마스크 쌍을 이용한 클래스 특화 시각 특징은 신장 종양 및 신장 낭종의 형태적·영상학적 특성을 효과적으로 반영하여 분할 성능을 향상시켰으며, 종양 크기 기반 support 영상 선택은 신장 종양과 신장 낭종 간의 오분할을 감소시키고 보다 안정적인 분할 결과를 제공하였다. 또한 신장 종양 분할에서 평균 성능 향상과 함께 성능 편차가 감소하는 경향을 확인하였으며 신장 종양 크기별 분석에서는 소형 종양과 극대형 종양에서 우수한 성능을 보여 제안 방법이 형태적 변이가 크거나 분할 난이도가 높은 사례에 효과적으로 적용될 수 있음을 확인하였다. 이러한 결과는 기존 CNN 및 Transformer 기반 방법과 비교하여, CLIP 기반 영상-언어 표현과 support 기반 시각 조건 정보를 결합한 제안 방법이 다양한 형태와 크기의 신장 종양에 대해 보다 강건한 특징 표현을 학습하고 안정적인 분할 성능을 제공할 수 있음을 보여준다.

향후에는 신장 종양 크기뿐 아니라 형태, 위치 및 영상 특징을 함께 고려하는 유사도 기반 support 검색 메커니즘을 도입하여 최적의 support 영상을 자동 선택하는 방향으로 확장할 예정이다. 또한 신장 종양과 신장 낭종 간의 공간적 관계를 활용하는 구조 인식(structure-aware) 조건 정보를 추가하여 오분할을 감소시키고, 다양한 장기 및 병변 분할 문제에 적용함으로써 제안 방법의 일반화 성능을 검증하고자 한다. 향후 연구에서는 최신 신장 종양 분할 기법을 포함한 다양한 최신 연구 방법과의 비교를 통해 제안 방법의 일반성과 성능을 보다 종합적으로 검증하고자 한다. 추가적으로 제안 방법을 RENAL nephrometry score 자동 산출 과정과 연계하여 신장 종양의 수술 난이도 평가 및 수술 계획 수립을 지원하는 임상 의사결정 보조 시스템으로 확장하고자 한다.

감사의 글

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원(HI22C1496) 받아 수행되었으며 이에 감사드립니다.

References

[1] S.A. Padala, A. Barsouk, K.C. Thandra, K. Saginala, A. Mohammed, A. Vakiti, P. Rawla, and A. Barsouk, "Epidemiology of Renal Cell Carcinoma," *World Journal of Oncology*, vol. 11, no. 3, pp. 79–87, 2020.

[2] P.V.A. Pinto, F.M.A. Coelho, A. Schuch, M. Zapparoli, and R.H. Baroni, "Pre-operative Imaging Evaluation of Renal Cell Carcinoma," *Revista da Associação Médica Brasileira*, vol. 70, no. Suppl 1, e2024S107, 2024.

[3] M.M. Picken, W. Lu, and N.G. Gupta, "Positive Surgical Margins in Renal Cell Carcinoma: Translating Tumor Biology Into Clinical Outcomes," *American Journal of Clinical Pathology*, vol. 143, no. 5, pp. 620–622, 2015.

[4] A. Kutikov, and R.G. Uzzo, "The RENAL nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth," *The Journal of Urology*, vol. 182, no. 3, pp. 844–853, 2009.

[5] S. Campbell, R.G. Uzzo, M.E. Allaf, et al., "Renal Mass and Localized Renal Cancer: AUA Guideline," *The Journal of Urology*, vol. 198, no. 3, pp. 520–529, 2017.

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

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *2016 Fourth International Conference on 3D Vision*, pp. 565–571, 2016.

[8] F. Isensee, P.F. Jaeger, S.A.A. Kohl, J. Petersen, and K.H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2021.

[10] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H.R. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," *International MICCAI Brainlesion Workshop*, pp. 272–284, 2021.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[12] Y. Ye, Y. Xie, J. Zhang, Z. Chen, and Y. Xia, "UniSeg: A Prompt-Driven Universal Segmentation Model as Well as A Strong Representation Learner," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer Nature Switzerland, Cham, vol. 14222, pp. 508–518, 2023.

[13] Z. Zhao, Y. Zhang, C. Wu, X. Zhang, X. Zhou, Y. Zhang, Y. Wang, and W. Xie, "Large-Vocabulary Segmentation for Medical Images with Text Prompts," *NPJ Digital Medicine*, vol. 8, no. 1, 566, 2025.

[14] W. Zhang, L. Luo, M. He, J. Hai, and J. Ye, "Organ-aware multi-scale medical image segmentation using text prompt engineering," *arXiv preprint*, arXiv:2503.13806, 2025.

[15] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B.A. Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21152–21164, 2023.

[16] E. Yang, C.K. Kim, Y. Guan, B.B. Koo, and J.H. Kim, "3D multi-scale residual fully convolutional neural network for segmentation of extremely large-sized kidney tumor," *Computer Methods and Programs in Biomedicine*, vol. 215, 106616, 2022.

[17] N. Heller, F. Isensee, K.H. Maier-Hein, X. Hou, C. Xie, F. Li, et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Medical Image Analysis*, vol. 67, 101821, 2021.

[18] N. Heller, F. Isensee, R. Tejpau, A. Wood, N. Papanikolopoulos, and C. Weight, "2023 Kidney and Kidney Tumor Segmentation Challenge," *Zenodo*, DOI: 10.5281/zenodo.7840134, 2023.

[19] KiTS23. [Online]. Available: https://github.com/neheller/kits23

〈저자소개〉

김 고 운

- 2023년 8월 서울여자대학교 소프트웨어융합학과 졸업(학사)
- 2026년 8월 서울여자대학교 컴퓨터학과 소프트웨어융합전공 졸업(석사)
- 관심분야 : 의료 인공지능, 딥러닝, 영상분할, 영상처리
- https://orcid.org/0009-0003-0694-2808

이 민 진

- 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

홍 헬 렌

- 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