Grounded-SAM-2 모델을 활용한 Zero-shot Annotation의 신뢰성 검증

이도현 [1], 최유경 [2]

[1] 세종대학교 스마트생명산업융합학과, [2] 세종대학교 AI 로봇학과

dhlee@sju.ac.kr, ykchoi@sejong.ac.kr

요 약

본 연구는 딥러닝 모델의 학습 데이터 구축 과정에서 발생하는 어노테이션 시간 및 비용 부담 경감을 위해, Zero-shot Learning 기반의 Grounded-SAM-2 모델이 어노테이션 보조 도구로서 활용될 수 있는 가능성을 검증하였다. 수직 재배 온실 내 참외 객체 분할 성능을 정량적·정성적으로 평가하였으며, 모델 성능에 영향을 미치는 입력 프롬프트, Box 및 Text 임계값, 사전 학습 가중치 등 핵심 파라미터들을 체계적으로 분석하였다. 특히 모델이 신뢰하는 예측의 품질을 보다 정확히 검증하고자 Matched Object mIoU 지표를 새롭게 도입하였다. 정량적 평가 결과, 모델은 BFV (Bed Full View) 이미지에서 Pixel-level mIoU 0.8037, Matched Object mIoU 0.8573 으로 우수한 분할 성능을 나타냈다. 반면, RFV (Robot Front Veiw) 이미지에서는 Pixel-level mIoU 0.5453 으로 상대적으로 낮은 성능을 보였으나, Matched Object mIoU 는 0.8333 을 기록하며 신뢰성 있는 예측에 대해서는 높은 분할 정확도를 유지함을 입증하였다. 정성적 분석에서는 작은 객체 탐지 및 겹친 객체 구분에서의 한계점이 관찰되었으나, 명확히 노출된 객체에 대해서는 우수한 품질을 보여 정량적 결과와 부합하였다. 이러한 결과들은 Grounded-SAM-2 모델이 수동 어노테이션을 완전히 대체하기는 어렵더라도, 신뢰성 있는 예측에 한하여 어노테이션 보조 도구로서 유의미하게 활용될 수 있음을 객관적으로 제시하는 데 기여한다. 연구의 재현성을 높이기 위해 본 연구에서 사용된 모든 코드는 https://github.com/editdiary/ZSL-LabelEase 에서 확인하실 수 있습니다.

1. 서론

딥러닝 모델의 성능은 양질의 대규모 데이터셋 확보에 크게 의존한다 [1]. 특히 이미지 분야에서 탐지(Detection), 분할(Segmentation)과 같은 고난이도 문제를 성공적으로 수행하기 위해서는 정교한 어노테이션(annotation) 작업을 통한 지도 학습 데이터 구축이 필수적이다. 그러나 이러한 어노테이션 작업은 막대한 시간과 비용을 요구하며, 이는 딥러닝 기술 적용의 주요 병목 현상으로 지적된다 [2].

이러한 문제를 해결하기 위해 다양한 자동화된 어노테이션 보조 도구들이 개발되어 왔다. 예를 들어, Roboflow 와 같은 상용 플랫폼은 Zero-shot 학습 모델을 활용한 'Smart Polygon' [3] 및 'Label Assist' [4] 기능으로 어노테이션 효율성 향상을 도모하고 있다. 하지만 이러한 도구들은 객체의 형태가 불분명하거나 배경이 복잡한 경우 성능이 저하되는 한계를 보이며, 특히 사전 학습 모델이 학습하지 않은 미지 클래스(unseen classes)에 적용이 어렵다는 제약이 있다. 또한, 기능 사용 횟수 제한 및 유료 결제 모델과 같은 상업적 제약도 존재한다.

오픈소스 어노테이션 도구인 Computer Vision Annotation Tool (CVAT)은 사용자가 직접 모델을 구축하여 자동 어노테이션 기능을 활용할 수 있는 유연성을 제공한다 [5]. 그럼에도 불구하고 CVAT 는 복잡한 설정 과정과 하드웨어 성능 요구치로 인해 높은 진입 장벽을 가지며, 복잡한 배경의 객체에 대한 성능 한계는 여전히 주요 해결 과제로 남아있다.

최근 Zero-shot Learning 분야의 발전은 이러한 한계를 극복할 잠재력을 보여주고 있다. 특히 Grounding DINO 는 텍스트 프롬프트 기반의 속성 비교를 통해 사전 학습되지 않은 객체에 대해서도 효율적인 탐지 능력을 입증했다 [6]. 이와 함께 Segment Anything Model (SAM)은 사용자 클릭과 같은 다양한 프롬프트에 반응하여 특정 객체를 정밀하게 분할하는 우수한 성능을 선보였다 [7]. 최근에는 Grounding DINO 와 SAM 의 강점을 결합한 Grounded-SAM 모델이 제안되어 더욱 정확하고 효율적인 객체 분할이 가능해졌다 [8].

이러한 혁신적인 Zero-shot 모델들은 수동 어노테이션을 대체할 강력한 보조 도구로서의 가능성을 제시하고 있다. 그러나 실제 복잡한 환경이나 특정 도메인에 특화된 객체에 대해 이러한 모델들이 얼마나 정확하고 신뢰성 있는 어노테이션 성능을 제

공하는지에 대한 체계적인 검증은 여전히 미흡한 실정이다. 본 논문에서는 Zero-shot Learning 기반의 Grounded-SAM-2 모델을 활용하여, 모델이 사전 학습하지 않은 객체인 참외(Korean melon, Chamoe)의 분할 성능을 정량적·정성적으로 평가하고자 한다. 이를 통해 Grounded-SAM-2 모델이 어노테이션에 필요한 노동력을 효과적으로 경감시키고 학습 데이터 구축의 효율성을 높이는 도구로서의 활용 가능성을 객관적으로 제시하고자 한다.

2. 재료 및 방법

본 장에서는 실험에 사용된 데이터셋 구성, 모델의 성능을 평가하기 위한 방법론, 그리고 실험이 진행된 상세 설정 및 컴퓨팅 환경에 대해 설명한다. 특히, Grounded-SAM-2 모델의 어노테이션 보조 도구로서의 실제 활용 가능성을 객관적으로 검증하고자, 기존 픽셀 단위 평가와 더불어 모델이 신뢰하는 예측의 품질을 중점으로 평가하는 새로운 지표를 도입하여 다각적인 분석을 수행하였다.

2.1 데이터셋

본 연구의 실험 데이터는 경상북도 성주군에 위치한 성주참외과채류연구소의 수직 재배 온실에서 수집되었다. 데이터 수집에는 DJI NEO 드론을 활용하였으며, 재배 공간의 특성을 고려하여 두 가지 유형의 영상을 촬영했다:

1. Bed Full View (BFV): 재배단 상단에서 드론의 틸트(tilt) 각도를 30 도 하향으로 설정하여 재배단 전체 영역을 조망하도록 촬영
2. Robot Front View (RFV): 로봇의 전방 시야를 모사하여, 재배단 사이의 통로를 따라 평행하게 전방을 향하도록 촬영

촬영된 영상은 1920x1080 해상도, 30 프레임으로 녹화되었고, 해당 영상으로부터 3 초 간격으로 프레임 이미지를 추출했다. 추출된 이미지 중 BFV 및 RFV 각 유형에서 무작위로 10 장씩 이미지를 선별하여 실험에 활용하였다.

그림 1. Bed Full View (BFV) 이미지 예시

그림 2. Robot Front View (RFV) 이미지 예시

선별된 이미지에 대해서는 CVAT 를 활용하여 수동 어노테이션을 수행함으로써 Ground Truth (GT) 분할 마스크를 생성했다. 어노테이션 대상은 온실 내 수확 가능한 성숙 참외로 한정되었다. 특히, 줄기나 잎과 같은 물체 의해 물리적으로 일부가 가려져 분리되어 보일지라도, 동일한 의미론적 객체로 판단되는 경우에는 이를 통합하여 어노테이션을 수행했다. 이는 객체 경계 상자(Bounding box)의 성능 비교를 명확하게 하기 위함이다. 그림 3 은 이러한 어노테이션의 예시를 보여준다.

그림 3. Bed Full View (BFV) 이미지 어노테이션 마스크 예시

2.2 성능 평가 방법

본 연구에서는 Grounded-SAM-2 모델의 어노테이션 보조 도구로서의 활용 가능성을 다각도로 검증하기 위해 크게 두 가지 유형의 성능 평가를 수행하였다.

1. 픽셀 단위 분할 정확도 평가:
   모델이 예측한 분할 결과와 GT 간의 픽셀 단위 분할 정확도를 확인하기 위해 Pixel-level mIoU (mean Intersection over Union)를 지표로 활용하였다. 이 지표는 이미지 전체 픽셀에 대한 모델의 전반적인 분할 성능을 정량적으로 나타낸다.
2. 매칭된 객체 분할 품질 평가:
   모델의 예측 결과가 특정 신뢰도 임계값(0.5)을 충족하여 GT 와 매칭된 객체에 대해 얼마

나 정확한 분할 품질을 제공하는지를 평가하고자 정밀도(Precision), 재현율(Recall), F1-score 및 Matched Object mIoU 를 지표로 활용하였다. 이러한 지표들은 모델이 스스로 신뢰할 수 있다고 판단하여 GT 와 성공적으로 매칭이 이루어진 예측 객체의 품질을 검증하는 데 중점을 둔다.

이를 위해 예측된 객체와 GT 객체 간의 일치 정도를 판단하는 새로운 지표인 Combined IoU 를 정의하였다. Combined IoU 는 객체의 위치 정보와 정확한 형태 정보를 동시에 고려하도록 Bounding Box IoU (bbox IoU)와 Mask IoU 를 3:7 의 비율로 가중 합산하여 계산된다. 이는 다음 수식 (1)과 같다.

$$\text{Combined IoU} = (0.3 \times \text{bbox IoU}) + (0.7 \times \text{Mask IoU}) \quad (1)$$

bbox IoU 는 객체의 대략적인 위치 판별에 유용하나, 정밀한 형태를 반영하기 어렵다. 반면 Mask IoU 는 객체의 세밀한 형태를 반영할 수 있지만, 픽셀 단위 평가의 특성상 모델의 예측이 GT 에서 미세하게 벗어나더라도 IoU 값이 크게 감소할 수 있는 한계가 있다. 따라서 Combined IoU 는 두 지표의 상호 보완적인 특성을 활용하여, 위치 정보와 분할 정보의 중요성을 동시에 반영하고 분할에 더 높은 가중치를 부여함으로써 정교한 판단을 가능하게 한다.

본 실험에서는 Combined IoU 가 임계값 0.5 이상인 경우에만 해당 예측 객체를 GT 객체와 성공적으로 매칭된 (True Positive) 것으로 간주하였다. Matched Object mIoU 는 이러한 기준을 통과하여 GT 와 매칭된 예측 객체들만을 대상으로 계산된 IoU 값의 평균을 의미한다. 이 지표는 모델의 전반적인 예측 성능이 낮더라도, 모델이 신뢰할 만하다고 판단한 예측에 한해서 높은 품질의 분할 마스크를 제공할 수 있는지 확인하기 위함이다. 궁극적으로 Matched Object mIoU 는 Grounded-SAM-2 모델이 자동 어노테이션 도구로서의 전면적인 대체는 아닐지라도, 어노테이션 보조 도구로서 '신뢰성 있는 예측'에 한하여 유의미하게 활용될 수 있음을 객관적으로 입증하는 데 기여한다.

2.3 실험 설정 및 컴퓨팅 환경

실험 과정에서는 객체 검출 및 분할 성능에 영향을 미치는 사전 학습 가중치, 입력 프롬프트, 그리고 바운딩 박스 임계값 및 텍스트 임계값 등 다양한 파라미터를 조정하며 실험을 진행하였다. 이러한 파라미터들이 모델의 최종 성능에 유의미한 영향을 미칠 수 있다고 판단하여, 최적의 파라미터 조합을 선정하고자 다각적인 시도를 수행했다.

본 연구에서 성능 검증에 활용된 컴퓨팅 환경은 표 1 에 상세히 제시되어 있다.

표 1. 컴퓨팅 환경 구성

Software		Hardware	
OS	Ubuntu 22.04 LTS	CPU	Intel Core I9-13900K
Python	3.10	GPU	NVIDIA RTX 4090
PyTorch	2.3.1	RAM	128GB
CUDA	12.1		

3. 실험 결과 및 분석

본 장에서는 Grounded-SAM-2 모델을 활용한 참외 객체 분할 실험의 정량적·정성적 분석 결과를 제시하고, 그 의미를 고찰한다. 모델의 핵심 파라미터 (입력 프롬프트, Box 및 Text 임계값)와 사전 학습 가중치가 성능에 미치는 영향을 체계적으로 분석하여, 어노테이션 보조 도구로서의 Grounded-SAM-2 모델의 활용 가능성 및 한계점을 다각도로 검증하였다. 모든 실험은 기준 모델 (Baseline) 설정을 기반으로 비교 분석되었다. 기준 모델은 Grounding DINO 의 Swin-T OGC 가중치와 SAM2 의 Hiera Large 가중치를 사용하였으며, Box 임계값 0.35, Text 임계값 0.25, 그리고 입력 프롬프트는 'yellow fruit'로 설정했다.

3.1 프롬프트 유형에 따른 성능 변화

Grounded-SAM-2 모델의 객체 분할 성능에 대한 입력 프롬프트의 영향을 평가하기 위해, 사전 학습 가중치, Box 및 Text 임계값을 비롯한 파라미터를 통제한 상태에서 다양한 프롬프트를 적용하여 실험을 진행했다. 그 결과는 표 2 에서 확인할 수 있다.

표 2. 입력 프롬프트에 따른 Pixel-level mIoU 성능 변화

Prompt	Pixel-level mIoU	
	BFV	RFV
yellow fruit (baseline)	0.7786	0.3050
yellow mature fruit	0.7970	0.5452
Korean melon	0.6878	0.4206
Chamoe	0.5905	0.2881
bright yellow Korean melon with white stripes	0.4142	0.2982
sweet yellow fruit with stripes also known as Korean melon	0.3449	0.1560

실험 결과, 입력 프롬프트는 모델의 성능에 직접적이고 유의미한 영향을 미치는 것으로 나타났다. 기준 프롬프트인 'yellow fruit'에 비해 'yellow mature fruit'를 입력했을 때 분할 성능이 미세하게 향상되었다. 특히 RFV 이미지에서는 0.2 이상의 상대적으로 큰 성능 증가를 보였는데, 이는 'mature'라는 표현이 참외의 성숙도와 같은 객체의 구체적인 시각적 특성을 모델에 더욱 효과적으로 전달했기 때문

으로 해석된다.

반면, ‘Korean melon’ 및 ‘Chamoe’와 같이 객체의 구체적인 명칭을 프롬프트로 사용했을 때는 오히려 성능이 감소했다. 이는 Grounded-SAM-2 와 같은 Zero-shot 모델이 ‘fruit’나 ‘yellow’와 같은 보편적인 개념에 비해 고유하거나 세부적인 개념을 추론하는 데 한계가 있음을 시사한다.

또한, ‘bright yellow Korean melon with white stripes’와 같이 객체의 특성을 문장 형태로 더욱 구체적으로 명시하는 프롬프트 역시 성능 감소를 야기했다. 이는 복잡하거나 불필요한 문맥 정보가 오히려 모델이 핵심적인 시각적 특징과 언어적 표현 간의 매핑을 정확히 수행하는 데 방해가 될 수 있음을 시사한다.

따라서 모델의 성능 향상을 위해서는 객체의 보편적이고 시각적으로 명확한 특징을 활용하여 간결하고 핵심적인 텍스트 프롬프트를 구성하는 것이 권장된다.

이후 실험에서는 입력 프롬프트 영향 분석을 통해 가장 우수한 성능을 보였던 ‘yellow mature fruit’를 기본 프롬프트로 채택하여 진행하였다.

3.2 Box 및 Text 임계값의 영향 분석

다음으로, Grounding DINO 모델의 Box 임계값이 모델 성능에 미치는 영향을 분석하고자 입력 프롬프트와 사전 학습 가중치를 고정한 채 실험을 수행하였다. 그 결과는 표 3 에 제시되어 있다.

표 3. Box 임계값 변화에 따른 Pixel-level mIoU 성능 변화

Box Threshold	Pixel-level mIoU	
	BFV	RFV
0.35 (baseline)	0.7970	0.5452
0.10	0.4358	0.0490
0.25	0.7650	0.2494
0.40	0.7913	0.5159
0.50	0.7512	0.4029

실험 결과, Box 임계값은 모델의 분할 성능에 상당한 영향을 미치는 것으로 나타났다. 그림 4 에서 볼 수 있듯이, 0.10 과 같이 낮은 임계값을 사용했을 때, 모델은 신뢰도가 낮은 예측 결과까지 과도하게 포함시키는 경향을 보였다. 이로 인해 온실 내 참외가 아닌 다른 객체(잎, 덕트 등)까지 불필요하게 분할 마스크에 포함되면서 전반적인 성능이 급격하게 감소하였다.

반면 0.50 과 같이 높은 임계값을 사용했을 때도 성능 감소가 관찰되었다. 이는 모델이 지나치게 보수적인 예측을 수행하여 신뢰도 높은 소수의 객체만을 분할하려 했기 때문으로 해석된다. 여러 실험을 통해 0.35 의 Box 임계값이 본 연구의 데이터셋에서 가장 최적의 성능을 보임을 확인하였으며, 이후 실험에서는 해당 값을 고정값으로 적용하였다.

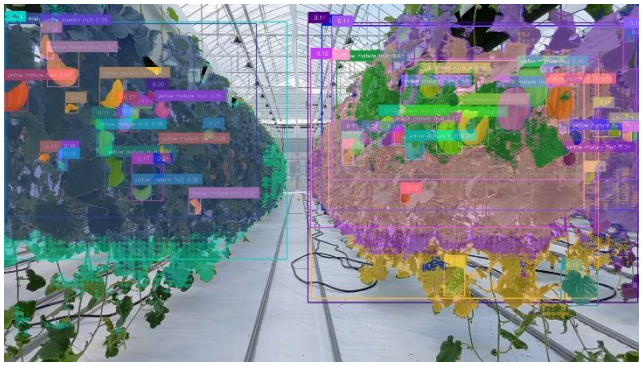

그림 4. 낮은 Box 임계값(0.10)에서의 과도한 객체 분할 예측 예시 (RFV)

한편, 텍스트 임계값은 본 실험에서 모델 성능에 유의미한 영향을 미치지 않는 것으로 나타났다. 0.05, 0.10, 0.25(Baseline), 0.50, 0.70 의 다양한 값으로 변경하며 성능을 테스트하였으나, 모든 경우에서 동일한 결과를 보였다. 이는 본 실험에서 사용된 프롬프트(yellow mature fruit)가 비교적 명확하고 간결하여 시각적 특징과의 매칭이 충분히 효과적으로 이루어졌음을 시사한다. 결과적으로 텍스트 임계값의 미세한 조정보다는 프롬프트가 지닌 정보의 본질적인 질이 모델 성능에 더욱 중요한 요인으로 작용함을 알 수 있다. 모든 값에서 성능 차이가 관찰되지 않았으므로, 이후 실험에서는 기준값인 0.25 를 텍스트 임계값으로 고정하였다.

3.3 사전 학습 가중치의 영향 분석

마지막으로 Grounding DINO 와 SAM2 모델의 사전 학습 가중치 변경이 Grounded-SAM-2 모델의 객체 분할 성능에 미치는 영향을 분석하였다. 그 결과는 표 4 에서 확인할 수 있다.

표 4. 사전 학습 가중치 변경에 따른 Pixel-level mIoU 성능 변화

Grounding DINO	SAM2.1	Pixel-level mIoU	
		BFV	RFV
SwinT_OGC	hiera_large	0.7970	0.5452
SwinB_cogcoor	hiera_large	0.6569	0.3526
SwinT_OGC	hiera_base+	0.8028	0.5442
SwinT_OGC	hiera_small	0.7969	0.5428
SwinT_OGC	hiera_tiny	0.8037	0.5453

실험 결과, Grounding DINO 모델의 사전 학습 가중치를 SwintT OGC 에서 더 복잡한 모델로 알려진 SwinB cogcoor 로 변경했을 때 오히려 성능이 감소하는 경향을 보였다. 한편 SAM2.1 의 가중치를 base 와 tiny 같이 더 간단한 모델로 변경할수록 성능이 미미하게 상승하는 경향을 나타냈다.

이러한 결과는 본 연구에서 사용된 참외 데이터셋의 특성과 관련이 있을 수 있다. 데이터셋의 복잡도가 상대적으로 더 단순하거나 특정 시각적 특징

이 두드러지지 않을 경우, 지나치게 복잡한 모델보다는 상대적으로 단순한 모델이 과적합을 덜 일으키고 더 효율적으로 일반화 성능을 발휘할 수 있음을 시사한다. 이는 모델 복잡도가 항상 성능 향상으로 이어지는 것은 아니며, 데이터셋의 특성에 맞는 모델 선택이 중요함을 보여준다.

또한, SAM2 의 사전 학습 모델 변경은 성능 차이가 미미했던 반면, Grounding DINO 의 가중치 변경은 성능 차이의 폭이 상대적으로 더 크게 나타났다. 이는 Grounded-SAM-2 시스템의 동작 방식과 연관 지어 해석할 수 있다. Grounded-SAM-2 는 Grounding DINO 에서 탐지한 바운딩 박스를 기반으로 SAM2 모델이 분할 마스크를 생성하는 구조이므로, 초기 객체 탐지 단계의 성능이 전체 시스템의 최종 분할 성능에 더욱 지배적인 영향을 미칠 수 있음을 의미한다. 즉, SAM2 의 분할 품질은 Grounding DINO 가 제공하는 바운딩 박스 정확도에 크게 의존하게 된다.

3.4 정량적 평가

앞선 실험들을 통해 선정된 최적 파라미터 조합의 Grounded-SAM-2 모델의 최종 정량적 예측 결과는 표 5 에 제시되어 있다.

표 5. 최적 모델의 BFV 및 RFV 평가 결과

	BFV	RFV
Precision	0.8082	0.6380
Recall	0.7171	0.3222
F1-score	0.7565	0.4208
Matched Object mIoU	0.8573	0.8333
Pixel-level mIoU	0.8037	0.5453

BFV (Bed Full View) 데이터셋에 대한 모델의 성능은 전반적으로 우수하게 나타났다. Pixel-level mIoU 가 0.8037 로 높은 수준을 기록하여, 픽셀 단위 분할 정확도가 뛰어남을 입증하였다. 더불어 Matched Object mIoU 는 0.8573 을 기록했으며, Precision, Recall, F1-score 또한 유사하게 높은 수치를 보였다. 이는 모델이 BFV 이미지에서 신뢰도 높은 예측 객체에 대해 정확한 분할 성능을 제공하며, 전반적인 픽셀 단위 예측 또한 성공적으로 수행함을 시사한다.

반면, RFV (Robot Front View) 이미지에 대한 모델의 성능은 BFV 이미지 대비 상대적으로 낮은 수치를 나타냈다. 특히 Pixel-level mIoU 는 0.5453 으로, BFV 이미지보다 현저히 낮은 수준을 기록하였다. Recall 과 F1-score 역시 낮은 수치를 보였는데, 이는 모델이 RFV 이미지 내의 실제 참외 객체들을 상당수 놓치고 있음을 (False Negative) 시사한다. 그럼에도 불구하고, Matched Object mIoU 는 0.8333 으로 비교적 높은 수준의 성능을 유지하였다. 이 결과는 모델의 전반적인 픽셀 단위 예측 성능이 낮고 재현율이 떨어지더라도, 모델이 신뢰할 만하다고 판단한 예측에 한해서는 높은 분할 정확도를 제공함을 의미한다.

3.5 정성적 평가

그림 5 와 그림 6 은 최적 파라미터 조합의 모델이 각 데이터셋 유형(BFV 및 RFV)에 대해 예측한 분할 결과를 보여준다.

그림 5. BFV 이미지에 대한 최적 모델의 예측 결과

BFV 이미지의 경우, 참외 객체가 상대적으로 명확하게 드러나 있어 모델이 대체로 우수한 분할 성능을 나타냈다. 그러나 다음과 같은 몇 가지 한계점들이 관찰되었다. 첫째, 매우 작은 크기의 객체에 대한 탐지율이 저조한 경향이 있었다. 둘째, 일부 이미지에서는 동일한 객체를 중복 탐지하는 현상이 관찰되었다. 셋째, 서로 겹쳐진 객체들을 정확히 구분하지 못하고 여러 객체를 하나의 단일 객체로 잘못 예측하는 경우가 있었다.

그림 6. RFV 이미지에 대한 최적 모델의 예측 결과

RFV 이미지에 대한 모델의 예측 성능은 BFV 이미지 대비 전반적으로 낮은 품질을 보였다. 이는 RFV 이미지가 BFV 이미지에 비해 객체의 크기가 상대적으로 작고, 부분 가림(occlusion) 현상이 두드러지며, 복잡한 배경으로 인한 시각적 난이도가 높기 때문으로 해석된다. 실제로 RFV 이미지에 대한 수동 어노테이션 과정에서도 객체의 모호성으로 인해 레이블링 난이도가 높았음을 고려하면, 이러한 결과는 예상 가능한 범위 내에 있다. 그럼에도 불구하고, 상대적으로 카메라에 가깝고 명확하게 노출된

객체에 한해서는 모델이 우수한 분할 성능을 유지하였다. 이러한 관찰 결과는 정량적 평가에서 RFV 데이터셋에 대한 Matched Object mIoU 가 0.8333 이라는 높은 수치를 기록한 것과 일치하며, 모델이 신뢰도 높은 예측에 대해서는 뛰어난 정확도를 유지함을 시사한다.

4. 결론

본 연구는 zero-shot learning 기반의 Grounded-SAM-2 모델을 활용하여 수직 재배 온실 환경 내 참외 객체에 대한 어노테이션 보조 도구로서의 활용 가능성을 탐색하고 그 성능을 정량적, 정성적으로 평가하였다. 다양한 입력 프롬프트, Box 및 Text 임계값, 그리고 사전 학습 가중치 변경 실험을 통해 모델의 성능 변화를 분석하였으며, 특히 Matched Object mIoU 라는 새로운 지표를 도입하여 모델이 신뢰하는 예측의 품질을 면밀히 검증하였다.

정량적 평가 결과, 모델은 BFV (Bed Full View) 이미지에서 Pixel-level mIoU 0.8037, Matched Object mIoU 0.8573 을 기록하며 우수한 분할 성능을 보였다. 반면 RFV (Robot Front View) 이미지에서는 Pixel-level mIoU 가 0.5453 으로 상대적으로 낮게 나타났는데, 이는 RFV 이미지의 복잡한 배경과 객체 가림 현상으로 인한 시각적 난이도 증가에 기인한 것으로 분석된다. 그럼에도 불구하고, RFV 이미지에서 Matched Object mIoU 가 0.8333 이라는 높은 수치를 기록했다는 점은 주목할 만하다. 이는 모델의 전반적인 예측 성능이 낮더라도, 스스로 신뢰할 만하다고 판단하여 매칭된 예측 결과에 한해서는 높은 수준의 분할 정확도를 제공함을 명확히 보여준다.

정성적 분석 또한 이러한 정량적 결과를 뒷받침한다. BFV 이미지에서는 전반적으로 좋은 성능을 보였으나 작은 객체 탐지 미흡, 중복 탐지, 겹쳐진 객체 구분의 한계점 등이 관찰되었다. RFV 이미지에서는 배경 복잡성으로 인한 낮은 재현율이 두드러졌음에도, 명확히 드러난 객체에 대해서는 뛰어난 품질을 유지하며 Matched Object mIoU 의 의미를 재확인시켜주었다.

종합적으로 판단했을 때, Grounded-SAM-2 모델은 수직 재배 온실 내 참외 객체 분할에 있어 예상보다 우수한 성능을 보였다. 특히 신뢰도 높은 객체 예측에 대해서는 두 이미지 유형 모두에서 0.8 이상의 높은 성능을 달성하여, 수동 어노테이션 작업을 부분적으로 대체할 수 있는 잠재력을 입증하였다. 이는 zero-shot 모델을 완전한 레이블링 대체 도구로 사용하기에는 신중해야 할 필요가 있으나, 모델이 정확하게 예측한 일부 결과에 한해서는 충분히 활용 가능한 보조 도구로서 노동 부담을 유의미하게 경감시킬 수 있음을 시사한다.

감사의 글

본 연구는 2025 학년도 1 학기 AI 로봇융합심화 PBL 농업 모듈 대학원 강의의 과제로서 수행되었습니다. 본 프로젝트를 통해 학습 및 발전에 많은 도움을 주신 최유경 교수님께 깊은 감사를 드립니다.

참고문헌

[1] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” arXiv preprint arXiv:1712.04621, 2017.

[2] CVAT, “Calculating the Cost Image Annotation for AI Projects,” CVAT Blog, 2024. [Online]. Available: https://www.cvat.ai/resources/blog/calculating-the-cost-of-solo-image-annotation-for-ai-projects. [Accessed: Jun. 22, 2025].

[3] “Smart Polygon.” Roboflow Docs. [Online]. Available: https://docs.roboflow.com/annotate/ai-labeling/enhanced-smart-polygon-with-sam. [Accessed: Jun. 22, 2025].

[4] “Label Assist.” Roboflow Docs. [Online]. Available: https://docs.roboflow.com/annotate/ai-labeling/model-assisted-labeling. [Accessed: Jun. 22, 2025].

[5] “Automatic annotation.” CVAT Documentation. [Online]. Available: https://docs.cvat.ai/docs/manual/advanced/automatic-annotation/. [Accessed: Jun. 22, 2025].

[6] S. Liu et al., “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” in *Proc. European Conference on Computer Vision*, Cham, Switzerland, 2024, pp. 38-55.

[7] N. Ravi et al., “SAM 2: Segment Anything in Images and Videos,” arXiv preprint arXiv:2408.00714, 2024.

[8] T. Ren et al., “Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks,” arXiv preprint arXiv:2401.14159, 2024.