강도다리 고밀도 수조의 영상 분할 AI 모델 비교 연구

김재영, 이래경, Mahnoor Ajmal*
한국전자통신연구원, *경북대학교
jaeyoung@etri.re.kr, laklee@etri.re..kr, *mahnoor.ajmal@knu.ac.kr

Comparative Analysis of AI Models for Image Segmentation in High-Density Spotted Halibut Tanks

Jae Young Kim, Lae Kyoung Lee, Mahnoor Ajmal*
Electronics and Telecommunications Research Institute, *Kyungpook Univ.

요 약

본 연구는 스마트 양식장에서 자동 급이를 위한 강도다리(Spotted halibut)의 중량 추정을 위해서 세 가지 대표적인 AI 기반 영상 분할 모델들(GroundingDINO+SAM, Mask R-CNN, Mask2Former)에 대해 비교·평가하였다. 수조를 촬영한 카메라 이미지들 중에 많은 강도다리들이 겹쳐 있어서 분할하기 어려운 대표적인 이미지들에 대해서 각 모델별 하이퍼파라미터를 조절하면서 영상 분할 결과를 비교하였다. 여러 개의 마스크 중에 과거 사육일지의 기간과 어류의 평균 체중 데이터를 기반으로 신뢰성이 있는 체중의 상·하한값을 추정하고 반영함으로써 강도다리를 특정하는 마스크를 선택하는 알고리즘을 개발하였다. 결론적으로 겹침이 적은 이미지에서는 세 모델 모두 안정적으로 분할되었으나, 겹침이 심한 이미지에서는 Mask2Former만이 일관되게 객체 분할이 가능하였다. 본 연구 결과는 스마트양식 시스템의 강도다리 생육 상태 모니터링, 자동 급이 제어 및 다양한 어류 관리 서비스 기능을 구현하는 데 활용될 수 있다.

Ⅰ. 서 론

국내 양식업은 사료비 비중이 높고 노동 의존도가 커서 생체량을 정확히 추정하고 급이량을 정밀 제어하는 능력이 곧 수익성과 직결된다. 스마트 양식장에서 어류 상태 관리 및 자동 급이를 안정적으로 수행하기 위해서 카메라 영상을 기반으로 개체 중량을 추정해 급이할 사료량을 정밀 산정할 필요가 있다. 그러나 강도다리(Spotted halibut)는 저면에 밀집·정지하는 시간이 길고 개체 겹침이 빈번해 단순 검출 기반 계수·중량 추정이 어렵다. 또한 카메라가 촬영한 강도다리 영상은 수면 반사·탁도·거품 등의 환경 잡음으로 인해 개별 개체 분할이 매우 어렵다. 본 연구는 이러한 문제들을 해결하기 위해서 대표적 객체 분할 접근 모델인 GroundingDINO 및 SAM(Segment Anything Model) 모델의 결합 방식 [3], [4], Mask R-CNN (2-stage 방식) [1], Mask2Former(masked attention을 이용한 마스크 분류 패러다임) [2] 을 동일 파이프라인에서 비교·평가하였다. 특히, 다양한 수조 영상 중 겹침이 심해 분할 난도가 높은 사례를 선별하고 각 모델의 하이퍼파라미터를 체계적으로 조절하며 결과를 비교하였다. 또한 과거 사육일지(기간, 평체)를 활용해 신뢰가능한 중량의 상·하한을 추정하고 그 구간에 부합하는 객체 만 선택하는 사육 데이터 기반 마스크 선택 알고리즘을 설계함으로써 개별 어류의 중량을 추정하는 데 신뢰성을 높였다. 본 연구는 결과적으로 생육 모니터링 - 중량 추정 - 급이 제어로 이어지는 스마트양식 시스템의 핵심 기반 기술을 제공할 수 있다.

Ⅱ. 본론

본 연구는 양식장 수조에 설치된 카메라로 촬영한 강도다리 영상 이미지를 입력받아 개별 어류를 분할하고, 마스킹한 면적을 기반으로 강도다리의 넓이와 중량을 추정할 수 있는 기법을 연구하였다. 데이터셋은 여러 마리가 겹쳐 하나의 덩어리로 보이는 난이도가 높은 이미지들을 대표 이미지로 선별해 모델의 한계를 나타날 수 있도록 설계하였다. 객체 분할 기법으로 GroundingDINO+SAM, Mask R-CNN, Mask2Former 모델들로 연구를 진행하였다. 먼저 GroundingDINO+SAM 파이프라인은 “fish, halibut”과 같은 텍스트 프롬프트로 바운딩 박스를 생성하고 [4], SAM에 바운딩 박스의 위치를 입력하여 정밀 마스크를 획득하는 방식이다 [3]. 이때 GroundingDINO의 텍스트-박스 매칭 점수 임계값 0.2, 텍스트 토큰 매칭 임계값 0.25로 하이퍼파라미터를 설정해서 결과를 도출하였다 [3], [4]. 강도다리의 겹침 정도가 작을 경우에는 마스크 생성이 잘 되었지만 상당이 많이 겹쳐져 있을 경우에는 거의 분할하지 못하는 결과가 나왔다. 그림 1은 강도다리 원본 이미지들과 GroundingDINO+SAM 모델을 적용해서 객체를 분할한 결과 이미지를 보여준다.

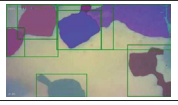
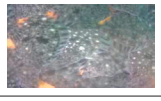
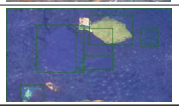
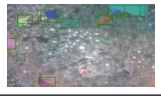
이미지 번호	1	2
원본 이미지		
Grounding+SAM 이미지		
이미지 번호	3	4
원본 이미지		
Grounding+SAM 이미지		

그림 1. 원본 이미지와 GroundingDINO+SAM 모델 적용 결과

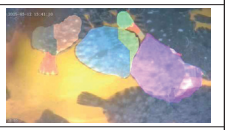
이미지 번호	1	2
Mask R-CNN 이미지		
이미지 번호	3	4
Mask R-CNN 이미지		

그림 2. Mask R-CNN 모델 적용 결과

둘째, Mask R-CNN은 COCO 사전학습 가중치를 출발점으로 강도다리 단일/소수 클래스 주석으로 도메인 미세학습을 수행하였다 [1]. 객체 인식 및 마스크 생성하는 모델 중에 그림 2와 같이 세 가지 모델들 중에 가장 좋지 못한 결과가 도출되었다. 셋째, Mask2Former는 마스크 어텐션을 이용한 마스크 분류 패러다임을 사용하며[2], Swin/ConvNeXt 백본과 결합해 상한 성능을 끌어올렸다. 그리고, 겹침이 심한 이미지에서 작은 마스크 필터링 및 워터셋 기반 겹침 분리 및 NUM_OBJECT_QUERIES(대략 200 - 300 범위), 최종 스코어 임계값 등과 같은 하이퍼파라미터를 설정하여 결과를 도출하였다. 세 개의 모델들 중에 분할 및 마스크 설정에서 가장 좋은 결과가 도출되었고 표 1과 같이 경량형 모델(Backbone :R50)과 대규모 모델(Backbone:Swin-L (IN21k))을 시험하였고 표 2와 그림 3과 같이 실행 결과를 도출하였다.

표 1. 경량형 모델 및 대규모 모델 비교

	Backbone	AP	weight	config
경량형 모델	R50	43.7	model_final_3c8ec9.pkl	maskformer2_R50_bs16_50ep.yaml
대규모 모델	Swin-L (IN21k)	50.1	model_final_e5f453.pkl	maskformer2_swin_large_IN21k_384_bs16_100ep.yaml

표 2. Mask2Former 시험 결과

	총 예측 객체 수	평균 지연시간(sec)	처리 속도(fps)
경량형 모델	92	4.488542258739471	0.22278948093959408
대규모 모델	41	3.468638598918915	0.2882975471447715

경량형 모델(R50)과 대규모 모델(Swin-L, IN21k) 구성을 함께 시험한 결과, 본 테스트셋과 파이프라인 설정에서 Swin-L 구성은 평균 지연시간이 더 낮고(≈3.47 s), 처리 속도는 더 높았으며(≈0.29 FPS), 예측 인스턴스 수는 더 보수적이었다. 이는 대규모 백본이 노이즈/과분할을 억제하면서도 고수준 표현으로 빠르게 수렴했기 때문으로 해석된다(단, 하드웨어·컴파일 옵션·입력 해상도·쿼리 수 등 구현 요인에 의해 달라질 수 있어 일반화에는 주의가 필요하다). 반대로 R50은 예측 인스턴스가 많고(≈92 vs. 41) 지연이 더 큰 양상을 보였는데, 이는 겹침 장면에서의 중복/오검출 및 후처리 비용 증가와 관련된 것으로 보인다.

Ⅲ. 결론

본 연구는 카메라 영상을 대상으로 객체 를 분할하고 마스크된 면적을 기반으로 강도다리의 넓이/중량 추정으로 이어지는 파이프라인을 구축하고, 겹침이 심한 강도다리 이미지에서 GroundingDINO+SAM, Mask R-CNN, Mask2Former의 성능을 비교·평가하였다. 데이터셋에서 난이도가 높은 대표 이미지를 선별하여 실험을 진행하였고 강도다리의 겹침 정

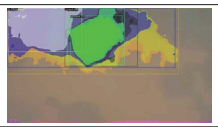
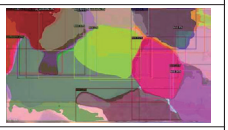
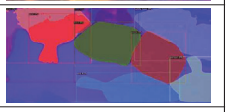
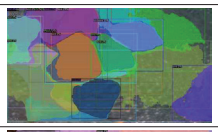
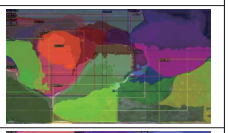
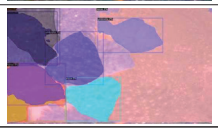
이미지 번호	1	2
경량형 모델		
대규모 모델		
이미지 번호	3	4
경량형 모델		
대규모 모델		

그림 3. Mask2Former 경량형 모델과 대규모 모델 적용 결과

도가 작을 때는 세 모델 모두 분할이 가능했으나, 겹침이 심한 장면에서 Mask2Former가 가장 안정적으로 객체를 분리하였다. GroundingDINO+SAM은 텍스트 프롬프트 기반 바운딩 박스 생성(텍스트-박스 매칭 0.2, 토큰 매칭 0.25)과 SAM 후처리로 간단히 적용 가능하다는 장점이 있으나, 다중 겹침 상황에서 바운딩 박스 단계의 모호성이 누적되어 분할 실패가 잦았다. Mask R-CNN은 소수 클래스 도메인 미세학습을 수행하였지만 고밀도·고겹침 이미지에서는 가장 성능이 좋지 않았다. 반면, Mask2Former는 마스크 어텐션 기반의 마스크 분류, 작은 마스크 제거·워터셋 기반 겹침 분리 및 점수 임계값 조정과 같은 후처리/하이퍼파라미터 세팅을 병행할 경우, 겹쳐 있는 객체들을 비교적 안정적으로 분리해냈다. 결론적으로 겹침이 빈번한 실제 양식장 영상에서는 Mask2Former가 가장 실용적인 선택이며, 향후 반지도학습·자기학습을 통해 Ground Truth 기법 및 최신 SOTA 모델들을 적용해서 완성도를 높일 예정이다.

ACKNOWLEDGMENT

본 연구 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음. [25ZD1150, 대경권 지역산업 기반 ICT 융합기술 고도화 지원사업(팜)]

참 고 문 헌

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” IEEE International Conference on Computer Vision (ICCV), 2017.

[2] B. Cheng, I. Misra, A. G. Schwing, and A. Kirillov, “Masked-Attention Mask Transformer for Universal Image Segmentation (Mask2Former),” CVPR, 2022.

[3] A. Kirillov, E. Mintun, N. Ravi, et al., “Segment Anything,” arXiv:2304.02643, 2023.

[4] S. Liu, Z. Zeng, T. Li, et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” arXiv:2303.05499, 2023.

[5] F. Meyer, “Topographic distance and watershed lines,” Signal Processing, 1994.