JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 10, pp. 269-280, October 2026
https://doi.org/10.9708/jksci.2026.31.10.269

Ontology-Guided Data Engine Framework for Defense Domain Data Construction

Seonghak Kim*

*Researcher, Agency for Defense Development, Daejeon, Korea

[Abstract]

The performance of defense Artificial Intelligence (AI) systems relies heavily on the acquisition of high-quality training data; however, constructing training datasets for rare military objects—such as enemy equipment and new weapon systems—poses significant challenges due to security constraints and the inherent scarcity of such data. To address this data scarcity problem, this paper proposes an Ontology-Guided Data Engine, an automated data construction framework that integrates the ontological structure of Wikidata with large vision foundation models. The proposed framework defines hierarchical concepts of military assets and utilizes visual prototypes from reference images to semantically filter valid data from the web, open datasets, and internal archives. Subsequently, it employs Grounding DINO and SAM to automatically generate precise bounding boxes and segmentation masks in the COCO format with minimal human intervention. Quantitative evaluations on 12 military classes demonstrate that the proposed engine achieves a data selection precision of 0.720 with a micro F1 of 0.729, an auto-labeling quality of 82.6% Box mAP and 83.9% Mask mAP, and the highest downstream detection and segmentation performance (54.9% Box mAP, 51.7% Mask mAP), outperforming keyword crawling, CLIP-only filtering, and SAM 3 baselines. These results confirm that the framework effectively identifies fine-grained differences in weapon systems that text-prompt-based models often fail to distinguish. Consequently, this framework is anticipated to serve as an efficient methodology for future research on Data-Centric AI within the defense domain.

▸ **Key words**: Defense AI, Data Scarcity, Data Mining, Ontology-Guided Data Engine, Auto-Labeling, Vision Foundation Model

• First Author: Seonghak Kim, Corresponding Author: Seonghak Kim
*Seonghak Kim (seonghak35@gmail.com), Agency for Defense Development
• Received: 2026. 04. 16, Revised: 2026. 08. 23, Accepted: 2026. 09. 28.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

[요 약]

국방 인공지능(AI) 시스템의 성능은 고품질 학습 데이터 확보에 크게 의존하지만, 적군 장비나 신규 무기 체계와 같은 희귀 국방 객체는 보안상의 제약과 원천적인 데이터 희소성으로 인해 학습 데이터셋 구축에 어려움이 따른다. 본 논문은 이러한 데이터 부족 문제를 해결하기 위해 위키데이터(Wikidata)의 온톨로지 구조와 비전 파운데이션 모델(Vision Foundation Model)을 결합한 자동화 데이터 구축 프레임워크인 온톨로지 기반 데이터 엔진(Ontology-Guided Data Engine)을 제안한다. 이는 군사 자산의 계층적 개념을 정의하고, 참조 이미지로부터 추출한 시각적 프로토타입(Visual Prototype)을 활용하여 웹, 공개 데이터셋 및 내부 아카이브로부터 유효한 데이터를 의미론적으로 필터링한다. 이후 Grounding DINO와 SAM을 적용하여 사람의 개입을 최소화하면서도 COCO 포맷의 정밀한 바운딩 박스(Bounding Box)와 분할 마스크(Segmentation Mask)를 자동으로 생성한다. 12개 군사 클래스를 대상으로 한 정량 평가에서 제안 엔진은 데이터 선별 정밀도 0.720(Micro F1 0.729), 오토 레이블링 품질 Box mAP 82.6% 및 Mask mAP 83.9%, 그리고 가장 뛰어난 다운스트림 탐지 및 분할 성능(Box mAP 54.9%, Mask mAP 51.7%)을 달성하여 키워드 크롤링, CLIP-only 필터링 및 SAM 3 베이스라인을 모두 상회하였다. 이는 본 프레임워크가 기존의 텍스트 프롬프트 기반 모델이 식별하기 어려운 무기 체계 간의 미세한 차이를 효과적으로 구분해 냄을 보여준다. 결론적으로, 본 연구는 향후 국방 분야의 데이터 중심 AI(Data-Centric AI) 연구를 위한 효율적인 방법론으로 활용될 것으로 기대된다.

▸ **주제어**: 국방 인공지능, 데이터 희소, 데이터 마이닝, 온톨로지 기반 데이터 엔진, 오토 레이블링, 비전 파운데이션 모델

I. Introduction

현대 전장에서 무인 정찰기(UAV), 고해상도 위성, 그리고 해상 감시 자산 등으로부터 수집되는 영상 정보의 양이 폭발적으로 증가함에 따라, 이를 자동으로 분석하는 지능형 감시 정찰(ISR) 시스템의 중요성이 대두되고 있다. 특히 딥러닝 기반의 객체 탐지(Object Detection)[1] 및 미세 식별(Fine-Grained Visual Recognition)[2], 그리고 객체의 정밀한 형상을 추출하는 인스턴스 분할(Instance Segmentation)[3] 기술은 전차, 군함, 전투기 등 전장의 핵심 자산을 식별하는 데 필수적이다.

그러나 국방 도메인은 일반적인 민간 분야와 달리, 고품질의 학습 데이터를 확보하는 데 있어 근본적인 어려움이 있다. MS-COCO[4]나 ImageNet[5]과 같은 대표적인 공개 데이터셋은 일상적인 객체 위주로 구성되어 있어 희소 군사 객체의 탐지 성능을 보장하지 못하며, 해양 및 항공 분야에서 제안된 대규모 데이터셋들 역시 민간 객체 중심이거나 일반적인 객체 범주(Generic Categories)로 레이블링(Labeling)되어 있어 전술적 가치가 제한적이다.

더욱이 군사 데이터는 보안 기밀 문제로 인해 수집 자체가 제한적이며, 전문가에 의한 데이터 레이블링은 막대한 비용과 시간이 소요되는 것이 현실이다. 특히 함정이나 항공기와 같은 복잡한 객체는 단순한 바운딩 박스(Bounding Box)만으로는 불충분하다. 이들 표적은 무장 발사대, 엔진 등 핵심 구성 요소를 정확히 구분하고, 위장(Camouflage)된 환경 속에서도 객체의 정확한 윤곽을 파악해야 하므로 픽셀 단위의 정교한 인스턴스 분할 정보가 필수적으로 요구된다. 하지만 이러한 고정밀 레이블을 수작업으로 구축하는 것은 현실적으로 불가능에 가깝다.

이러한 데이터 희소 문제를 해결하기 위해, 본 논문에서는 공개된 웹 데이터와 공개 데이터셋, 그리고 기관이 자체적으로 보유하고 있는 방대한 내부 데이터를 통합적으로 활용하여 군사 도메인에 특화된 학습 데이터를 구축하는 '온톨로지 기반 데이터 엔진(Ontology-Guided Data Engine)' 프레임워크를 제안한다. 본 프레임워크는 외부 데이터뿐만 아니라, 레이블링 비용 문제로 활용되지 못하고 있던 기관 내부의 비정형 영상 데이터에서도 사용자가 원하는 특정 객체를 자동으로 추출하고 레이블링할 수 있도록 설계되었다.

본 프레임워크는 크게 두 단계로 구성된다. 첫째, 위키데이터(Wikidata)[6] 지식 그래프를 통해 군사 객체의 계층적 개념(Hierarchical Concepts)을 정의하고, 이를 기

www.kci.go.kr

준으로 삼아 내·외부의 방대한 데이터 속에서 유의미한 표적 이미지만을 의미론적으로 필터링(Semantic Filtering)한다. 둘째, 선별된 데이터에 대해 Grounding DINO[7]와 SAM[8]과 같은 최신 비전 파운데이션 모델(Vision Foundation Model)을 결합하여, 추가적인 수동 레이블링 없이도 정교한 바운딩 박스와 분할 마스크(Segmentation Mask)를 생성하는 파이프라인을 수행한다. 특히 이 파이프라인은 보안이 필수적인 폐쇄망 환경에서도 독립적으로 구동 가능하며, 데이터 희소성이 높은 희귀 클래스까지 체계적으로 수집할 수 있다.

제안된 프레임워크의 효용성을 검증하기 위해 최신 비전 파운데이션 모델인 SAM 3[9]와의 정성적·정량적 비교 평가를 수행하였다. 실험 결과, 텍스트 프롬프트에 의존하는 기존 모델이 '탄도탄 요격 미사일'과 '함대공 미사일' 같이 시각적 형상은 유사하나 전술적 용도가 다른 객체를 혼동하는 것과 달리, 본 프레임워크는 온톨로지 기반의 시각적 프로토타입을 통해 이를 명확히 구분해내었다. 또한, '지뢰 제거 전차'와 같은 희귀 클래스에 대해서도, 기존 모델이 겪는 미검출이나 일반적인 '보병 전차'로의 오인 문제를 극복하고 정확한 식별 및 레이블링이 가능함을 확인하였다.

II. Related Works

1. Defense Domain Datasets

범용 객체 탐지 모델은 MS-COCO[4]와 같은 대규모 공개 데이터셋을 기반으로 발전해왔으나, 국방 도메인에서는 보안 제약으로 인해 실제 데이터의 수집과 공개가 근본적으로 제한된다. 해양 분야의 MARVEL[13]은 200만 장 이상의 선박 이미지를 포함하지만 레이블이 단순하고 노이즈가 많으며, SeaShips[14]는 민간 상선 위주로 구성되어 있다. 항공 및 위성 영상 분야에서는 DOTA[10], VisDrone[11], xView[12] 등 대규모 데이터셋이 공개되어 있으나, 이들은 민간 원격탐사 객체 중심으로 '비행기'와 같은 일반적인 클래스만 포함되어 탄도탄 요격 미사일과 같은 세분화된 무기 식별에는 활용이 어렵다. 이러한 데이터 공백을 메우기 위해 시뮬레이션 기반 합성 데이터 생성 연구가 시도되어 왔으나, 합성-실사 간 도메인 격차(Domain Gap)와 3D 자산 제작 비용이 새로운 병목이 된다. 본 연구는 이러한 한계를 극복하기 위해 웹 또는 내부 아카이브에 존재하는 데이터를 파이프라인 전반에 걸쳐 자동으로 수집, 정제, 레이블링함으로써, 도메인 격차를 최소화하는 고품질 데이터셋을 최소 비용과 인력만으로 안정적으로 구축한다.

2. Ontology-based Data Mining

지식 구조를 시각 인식에 결합하려는 연구는 웹 이미지로부터 시각 개념을 축적하는 NEIL[15], 웹 검색으로 개념의 모든 변형을 학습하는 Webly-supervised 접근[16], 지식 그래프의 관계 구조를 그래프 신경망으로 전파하여 제로샷 분류에 활용하는 연구[17] 등으로 진행되어 왔다. 공개 지식 그래프인 Wikidata[6]는 1억 개 이상 개체의 계층 관계와 대표 이미지를 구조화된 형태로 제공하여 이러한 연구의 기반이 되고 있지만, 기존 연구들은 지식 구조를 주로 분류기의 추론 보조나 클래스 간 관계 표현에 활용하였으며, 데이터가 희소한 도메인에서 수집 대상 클래스 자체를 계층적으로 정의하고 참조 이미지 확보까지 이끄는 수집 프로세스의 프레임워크로 사용한 사례는 드물다. 본 연구는 위키데이터 및 위키미디어 온톨로지를 국방 도메인 특화 세부 개념의 계층적 정의 및 하위 개념별 참조 이미지 자동 확보를 이끄는 핵심 파이프라인으로 활용하여 수집의 자동화 범위를 확장한다.

3. Auto-Labeling with Vision Foundation Model

대규모 이미지-텍스트 쌍으로 사전 학습된 CLIP[18]과 데이터 큐레이션을 개선한 MetaCLIP[19]은 임의의 텍스트와 이미지를 공통 임베딩 공간에서 비교할 수 있게 하여 개방 어휘(Open-vocabulary) 인식의 기반을 마련하였다. 객체 수준에서는 텍스트 프롬프트로 임의 클래스의 바운딩 박스를 추정하는 Grounding DINO[7]와 프롬프트 기반 분할 모델 SAM[8]이 등장하였고, 두 모델을 결합한 Grounded SAM[20]이나 탐지와 분할을 단일 모델로 통합하고 텍스트 및 예시(Exemplar) 프롬프트를 지원하는 SAM 3[9]와 같이 오토 레이블링 파이프라인으로 확장되었다. 그러나 이들 모델은 공통적으로 텍스트 프롬프트의 표현력에 의존하므로, 언어로는 구분이 모호하지만 시각적으로는 구별되는 미세 군사 클래스에서 성능이 저하되는 한계가 있다. 본 연구는 단순 텍스트 프롬프트를 넘어 참조 이미지로부터 유도된 고차원 시각 프로토타입(Visual Prototype)을 결합하여 필터링 및 레이블링 정밀도를 극대화하며, Grounding DINO와 SAM을 연계해 최소한의 사람 검수 비용만으로 다운스트림 탐지 및 분할 모델 학습에 활용 가능한 고품질 레이블링 데이터셋을 안정적으로 생성한다.

www.kci.go.kr

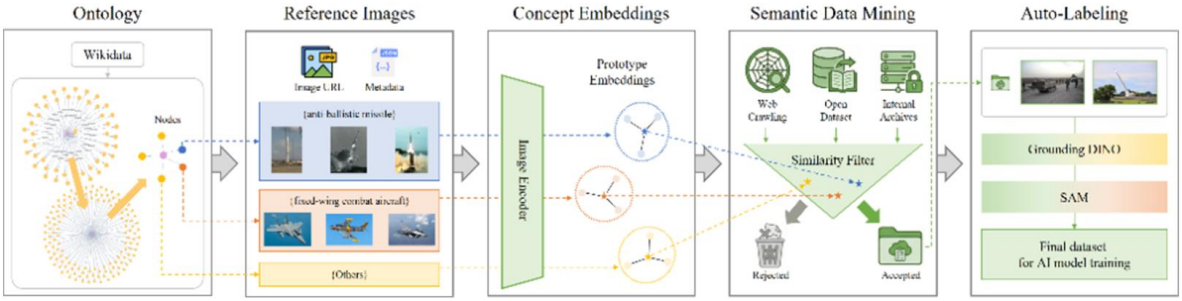


Fig. 1. Overview of the ontology-guided data engine framework

III. Data Engine Framework

본 연구에서 제안하는 온톨로지 기반 데이터 엔진(Ontology-Guided Data Engine)은 데이터 정의부터 수집, 정제, 레이블링까지의 전 과정을 자동화한 파이프라인이다. Fig. 1에 도식화된 바와 같이, 해당 시스템은 (1) 온톨로지 구축, (2) 참조 이미지 수집, (3) 개념 임베딩 생성, (4) 의미론적 데이터 마이닝, (5) 파운데이션 모델 기반 오토 레이블링의 5단계 워크플로우로 구성된다.

1. Ontology

데이터 엔진의 첫 단계는 수집 대상인 군사 객체의 개념적 범위를 정의하고, 이들 간의 계층적 관계를 수립하는 것이다. 본 연구는 방대한 지식 그래프인 위키데이터(Wikidata)를 활용하여 이를 자동화한다.

Fig. 2는 이 과정을 시각적으로 보여준다. 사용자가 입력한 최상위 개념(Root Concept)을 시작점으로 하여, SPARQL 쿼리[21]를 통해 하위 개념을 재귀적으로 탐색한다. 예를 들어, 미사일(Missile)을 최상위 개념으로 설정할 경우, 데이터 엔진은 이를 지대공 미사일(Surface-to-air missile)과 같은 중간 카테고리로 확장하고, 탄도탄 요격 미사일(Anti-ballistic missile)과 같은 구체적인 하위 개념에 도달하는 정교한 온톨로지 구조를 생성한다. 이 과정은 수작업 분류 체계가 가질 수 있는 주관적 편향을 배제하고, 객관적이고 확장 가능한 분류 체계를 제공한다.

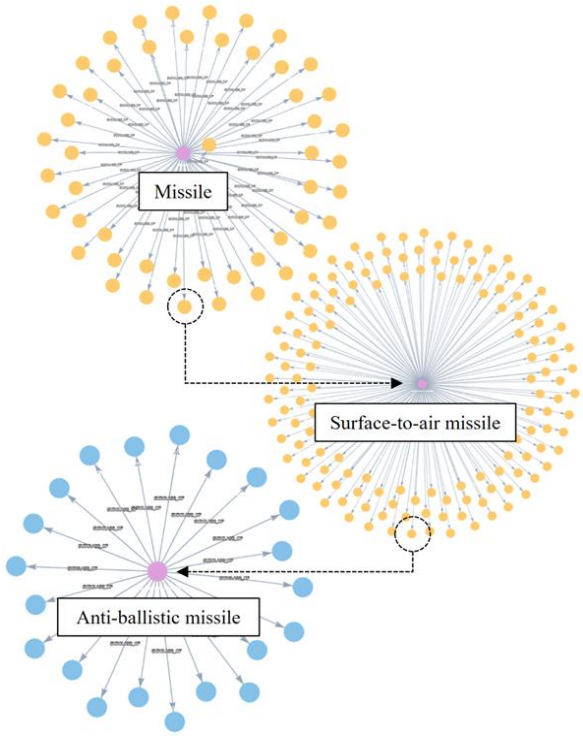


Fig. 2. Ontology-based concept definition and expansion process

2. Reference Images

앞서 구축된 온톨로지의 각 노드에 해당하는 개념에 대해, 위키미디어 공용(Wikimedia Commons)으로부터 검증된 이미지를 수집하여 참조 이미지를 구성한다. 위키미디어의 데이터는 불특정 다수의 웹 크롤링 데이터와 달리, 검증된 메타데이터와 명시적인 의미론적 연결을 기반으로 관리되므로, 레이블 노이즈(Label Noise)가 최소화된 고신뢰도(High-fidelity) 참조 데이터를 제공한다. 따라서 이 참조 이미지 세트는 해당 클래스의 시각적 특징을 가장 전형적으로 보여주는 기준 데이터로서, 후속 단계인 의미론적 마이닝의 정확도를 보장하는 역할을 수행한다.

www.kci.go.kr

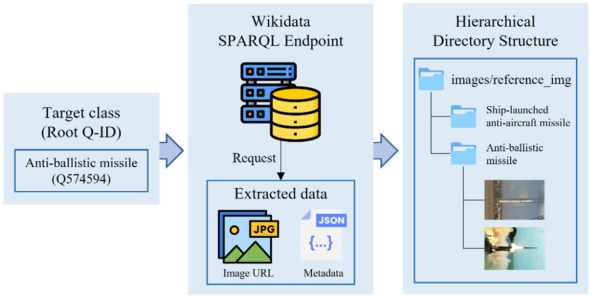


Fig. 3. Reference image collection pipeline

Fig. 3은 이러한 참조 이미지 수집 파이프라인을 보여준다. 시스템은 앞서 획득한 타겟 클래스의 루트 Q-ID를 기준으로 위키데이터 SPARQL 엔드포인트에 질의한다. 쿼리는 해당 개념의 하위 항목들에 대해 이미지 속성의 URL과 관련 메타데이터를 요청한다. 이후 추출된 정보는 코드에 정의된 파일명 규칙에 따라 처리되어, 로컬 스토리지의 계층적 디렉토리 구조에 자동으로 저장된다. 이러한 체계적인 저장 방식은 추후 임베딩 생성 단계에서 각 파일의 출처를 명확히 추적할 수 있게 한다.

3. Concept Embeddings

확보된 참조 이미지 세트는 각 클래스의 시각적 지식(Visual Knowledge)을 집약하고 있다. 본 단계에서는 이들을 단일한 고차원 벡터로 압축하여, 해당 클래스를 대표하는 프로토타입 임베딩(Prototype Embedding)을 생성한다. 구체적으로, 거대 비전-언어 모델인 MetaCLIP[19]을 백본 인코더(Backbone Encoder)로 활용하며, open_clip 라이브러리의 ViT-H-14 아키텍처에 MetaCLIP 2.5B 데이터셋으로 사전 학습된 metaclip_fullcc 체크포인트를 사용한다. Fig. 4는 이러한 프로토타입 생성 과정을 시각적으로 보여준다.

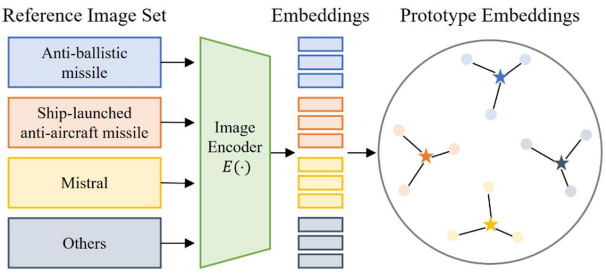


Fig. 4. Prototype embedding generation process

특정 클래스 C에 속한 N개의 참조 이미지 세트 $I_C = \{I_1, I_2, \cdots, I_N\}$에 대해, 사전 학습된 인코더 E를 통과시켜 얻은 개별 특징 벡터들을 평균하여 클래스 프로토타입 임베딩 P_C을 산출한다.

$$P_C = \frac{1}{N}\sum_{i=1}^{N}\frac{E(I_i)}{\| E(I_i) \|}$$

이러한 평균화 방식은 개별 참조 이미지에 포함될 수 있는 배경 잡음(Background Noise)이나 특이값(Outlier)을 희석시키고, 클래스 고유의 전형적인 시각 특징만을 남기는 효과가 있다. 특히, 이는 단순 텍스트 프롬프트(Text Prompt)만으로는 묘사하기 어려운 미세한 형상 차이를 벡터 공간 내에 함축적으로 보존한다. 결과적으로 생성된 프로토타입 임베딩은 데이터 마이닝 단계에서 텍스트 검색의 모호성을 보완하고, 수집된 데이터의 시각적 일치성을 검증하는 강력한 기준점이 된다.

4. Semantic Data Mining

본 단계에서는 개념 임베딩을 통해서 생성된 클래스별 프로토타입 임베딩 P_C을 기준점으로 하여, 다양한 출처의 이질적인 데이터 소스(Heterogeneous Sources)로부터 유의미한 학습 데이터를 선별한다. 데이터 마이닝의 대상은 크게 세 가지로 분류될 수 있다: (1) 웹 검색 이미지(Web Crawling), (2) 기존 공개 데이터셋(Open Datasets), 그리고 (3) 기관 내부의 비정형 데이터 아카이브(Internal Raw Archives)이다. 특히, 레이블링 비용과 보안 문제로 인해 활용되지 못하고 방치되었던 내부 영상 자산을 학습 가능한 형태로 자산화할 수 있다는 점에서 중요한 의의를 갖는다.

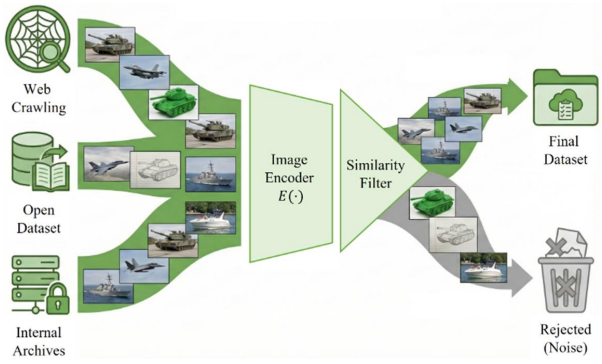


Fig. 5. Semantic data mining and filtering process

구체적인 마이닝 프로세스는 다음과 같다. 먼저, 텍스트 쿼리나 전체 순회(Full Scan)를 통해 대규모 후보 이미지 집합 $D = \{I_1, I_2, \cdots, I_M\}$을 확보한다. 이후, 각 후보 이미지 I_k를 동일한 인코더 E로 임베딩한 후, 해당 클래스의 프로

www.kci.go.kr

토타입 P_C과의 코사인 유사도(Cosine Similarity)를 계산한다. 이 유사도 점수가 특정 임계값(Threshold) 이상인 이미지만을 최종 데이터셋에 포함시키고, 나머지 이미지는 제거한다. 유사도 임계값은 수집 데이터 풀의 특성에 따라 최적값이 달라지므로 단일 상수로 고정하지 않고, 성능 평가에 사용되지 않는 별도의 소규모 검증용 데이터셋에서 Micro F1을 최대화하는 값으로 설정한다. 본 연구의 실험에서는 이 절차에 따라 0.85를 적용하였다.

$$CS(I_k, P_C) = \frac{E(I_k) \cdot P_C}{\| E(I_k) \| \, \| P_C \|}$$

Fig. 5는 이러한 의미론적 필터링 과정을 시각적으로 보여준다. 이 과정은 텍스트 키워드 크롤링(Keyword Crawling) 방식이 필연적으로 수반하는 레이블 노이즈(Label Noise)를 획기적으로 제거한다. 예를 들어, 'Tank'라는 키워드로 수집된 이미지 중 '물탱크(Water tank)', '장난감 전차', '게임 스크린샷' 등 시각적 특징이 실제 군사 객체와 다른 데이터들은 프로토타입 벡터와 유사도가 낮게 산출되어 자동으로 배제된다. 결과적으로, 본 프레임워크는 사람의 검수를 최소화하면서도 데이터의 품질을 보장하며, 다양한 소스로부터 일관된 품질의 이미지를 획득할 수 있게 한다.

5. Auto-Labeling

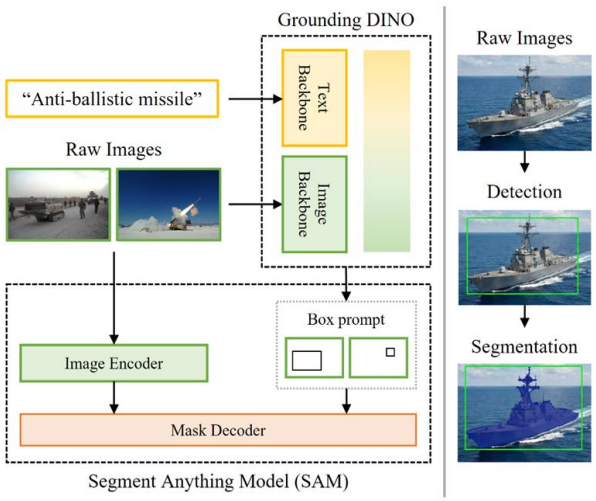


Fig. 6. Foundation model-based auto-labeling pipeline

의미론적 마이닝을 통해 수집된 데이터는 레이블이 없는 원본 이미지(Raw Image) 상태이다. 본 단계에서는 최신 비전 파운데이션 모델들의 제로샷(Zero-shot) 추론 능력을 결합하여, 사람의 개입 없이도 정교한 바운딩 박스(Bounding Box)와 분할 마스크(Segmentation Mask)를 생성하는 자동화된 레이블링 파이프라인을 수행한다. Fig. 6은 이러한 오토 레이블링 파이프라인의 과정을 보여준다.

5.1 Open-Set Object Detection

먼저 Grounding DINO[7] 모델을 활용하여 객체 위치를 추정한다. 온톨로지에서 정의된 클래스 명(예: '탄도탄 요격 미사일')을 텍스트 프롬프트(Text Prompt)로 입력하면, 모델은 이미지 내에서 해당 설명과 의미론적으로 일치하는 객체를 찾아 바운딩 박스를 예측한다. 이 과정에서 신뢰도 점수(Confidence Score)가 설정된 임계값(Threshold) 미만인 예측은 노이즈로 간주하여 필터링하고, 비최대 억제(NMS, Non-Maximum Suppression) 알고리즘[22]을 적용하여 중복된 박스를 제거함으로써 탐지의 정확도를 높인다. 본 연구에서는 박스 신뢰도 임계값 0.35, 텍스트 임계값 0.25, NMS IoU 임계값 0.50을 적용하였다.

5.2 Instance Segmentation

검출된 바운딩 박스는 SAM[8]의 입력 프롬프트(Box Prompt)로 전달된다. SAM은 박스 내부의 시각적 정보를 분석하여 객체의 경계를 픽셀 단위로 정밀하게 분할한 마스크를 생성한다. 이는 단순한 사각형 영역을 넘어, 항공기의 날개 끝부분 같은 불규칙하고 복잡한 형상까지 정확하게 식별해낼 수 있게 한다. 본 연구에서는 SAM ViT-H 체크포인트를 사용하였다.

최종적으로 생성된 레이블 데이터는 표준 COCO 포맷[4]으로 변환되어 저장되며, 이를 통해 별도의 추가 가공 없이도 다양한 딥러닝 모델의 학습 및 평가에 즉시 활용 가능한 고품질의 데이터셋이 완성된다.

IV. Experiments

1. Pilot Data Construction

본 연구에서 제안하는 온톨로지 기반 데이터 엔진 프레임워크의 성능을 검증하기 위해, 자체 구축 온톨로지에서 '미사일(Missile)'을 루트 노드(Root Node)로 설정하여 시범 데이터를 구축하였다.

Fig. 2에 나타난 바와 같이, 데이터 엔진은 '미사일' 하위에 존재하는 61개의 중간 카테고리(예: 'ship-launched missile', 'anti-satellite missile', 'surface-to-air

www.kci.go.kr

missile' 등)를 자동으로 식별하고 확장하였다. 나아가 각 중간 카테고리는 다시 세부적인 하위 기종으로 분류되는데, 예를 들어 'surface-to-air missile' 카테고리 하위에는 'anti-ballistic missile'이 포함되어 있고, 여기에는 'L-SAM', 'M-SAM', 'LIM-49A Spartan' 등 21개의 구체적인 리프 노드(Leaf Nodes)가 존재함을 확인하였다. 또 다른 예시 상위 개념들의 계층적 통계는 Table 1에서 확인할 수 있다.

이러한 계층적 접근은 단순히 다양한 미사일을 'Missile'이라는 단일 범주로 통합하던 기존 방식과 달리, 전술적 용도와 플랫폼에 따라 세부 클래스를 정교하게 구분함으로써 고성능 식별 모델 개발을 위한 최적의 데이터셋을 제공한다.

또한, 텍스트 검색만으로는 데이터 확보가 어려운 희귀 클래스에 대해서도, 본 데이터 엔진은 위키미디어 참조 이미지를 시드(Seed)로 활용한 의미론적 마이닝을 통해 클래스당 평균 수백 장 이상의 유효 데이터를 확보할 수 있게 해준다. 이는 본 프레임워크가 인터넷 웹뿐만 아니라 공개된 범용 데이터셋 및 내부 아카이브에 산재한 희소 군사 자산을 발굴하여 자산화하는 데 매우 효과적임을 시사한다.

Table 1. Hierarchical statistics

Root Node	Mid Category	# Variant
Missile	anti-radiation missile	22
	anti-aircraft missile	5
	anti-submarine missile	18
	anti-satellite missile	1
	surface-to-air missile	150
	... (61 total)	...
Military aircraft	fixed-wing military aircraft	4
	electronic warfare aircraft	27
	flying bomb	1
	... (688 total)	...
Warship	capital ship	1
	turret ship	2
	aircraft carrier	51
	... (72 total)	...
Tank	demining tank	1
	infantry tank	7
	... (217 total)	...

2. Experimental Setup

제안 프레임워크의 성능을 정량적으로 검증하기 위해 12개 군사 클래스에 대해 100장씩 총 1,200장의 후보 이미지를 수집하고, 전체 데이터에 대해 사람이 직접 목표 클래스와의 관련성을 검수하여 데이터 선별 품질 평가를 위한 정답(Ground Truth)을 확정하였다. (1) 키워드 크롤링, (2) CLIP-only 필터링, (3) 제안하는 데이터 엔진의 선별 결과를 정답과 비교하여 정밀도(Precision), 재현율(Recall), Micro F1, 정확도(Accuracy)를 측정하였다.

평가 대상으로 선정된 12개 클래스는 Table 1에 명시된 다수의 하위 범주 중, 온톨로지의 최상위 4개 계열(Missile, Military aircraft, Warship, Tank)에서 각각 3개씩 선정하였으며, 각 계열 내에서는 객체 크기, 형태, 웹 수집 난이도가 다양하도록 구성하였다.

또한 레이블링 품질 및 다운스트림 성능 평가를 위해, 사람이 직접 검수한 바운딩 박스 및 분할 마스크 레이블을 갖는 250장의 국방 도메인 테스트셋을 별도로 구축하였다. 각 방법이 선별한 데이터셋을 학습 데이터로 활용하여 YOLOv8s-seg[23] 모델을 배치 크기 16, 입력 이미지 크기 640, 100 에폭의 동일 조건으로 학습시킨 후, 250장 테스트셋에서 Box mAP@50:95 및 Mask mAP@50:95를 측정하였다. 이때 각 방법이 선별한 클래스 비율을 유지하는 층화 추출(Stratified Sampling)을 통해 세 방법의 학습 데이터 규모를 동일하게 통제하여, 학습 데이터의 양이 아닌 선별 품질의 차이만이 성능에 반영되도록 하였다. 전체 실험 및 측정은 단일 NVIDIA RTX 3090 GPU 환경에서 수행되었다.

본 데이터 엔진 내부에서 각 파운데이션 모델은 단계별로 서로 다른 역할을 수행한다. MetaCLIP[19]은 이미지와 텍스트를 공통 임베딩 공간으로 사상하는 인코더로서 시각 프로토타입 생성과 의미론적 필터링을 담당하고, Grounding DINO[7]는 필터링을 거친 이미지에서 객체의 바운딩 박스를 추정하는 탐지를, SAM[8]은 경계 상자를 박스 프롬프트로 전달받아 픽셀 단위 분할 마스크를 생성한다. 반면 비교 대상인 SAM 3[9]는 텍스트나 예시(Exemplar) 프롬프트를 입력받아 탐지와 분할을 단일 모델로 수행하는 통합 파운데이션 모델로서 전체 파이프라인의 성능을 검증하기 위한 베이스라인으로 활용되었다.

www.kci.go.kr

3. Quantitative Evaluation

Table 2. Data selection quality (top) and per-class F1 distribution (bottom)

Method	Precision	Recall	F1	Acc.
Keyword crawling	0.518	1.000	0.682	0.518
CLIP-only filtering	0.560	0.861	0.679	0.578
Data Engine	**0.720**	**0.738**	**0.729**	**0.716**

Method	Per-class F1 std (min-max)
Keyword crawling	0.164 (0.431-0.926)
CLIP-only filtering	0.184 (0.273-0.919)
Data Engine	**0.147 (0.433-0.936)**

Table 2는 데이터 선별 품질의 정량 평가 결과를 보여준다. 키워드 크롤링은 후보 이미지를 전혀 버리지 않아 재현율이 1.000이지만 정밀도가 0.518에 불과하여, 수집된 데이터의 절반가량이 목표 클래스와 무관한 노이즈임을 보여준다. CLIP-only 필터링 역시 정밀도 0.560으로 키워드 크롤링과 큰 차이를 보이지 못한 반면, 데이터 엔진은 정밀도 0.720, Micro F1 0.729를 기록하여 불필요한 데이터를 효과적으로 제거하고 목표 클래스에 부합하는 이미지를 선별하였다.

또한, 클래스별 F1 분포를 보면, 데이터 엔진은 표준편차 0.147, 최소 0.433(infantry tank), 최대 0.936(aircraft carrier)로, 키워드 크롤링의 표준편차 0.164 및 CLIP-only 필터링의 표준편차 0.184에 비해 클래스 간 편차가 작았다. 이는 제안 방법의 성능 개선이 특정 클래스에 편중된 것이 아니라 클래스 전반에 걸쳐 고르게 나타남을 의미한다.

Table 3. Downstream detection and segmentation performance

Method	Box mAP@50:95	Mask mAP@50:95
Keyword crawling	0.493	0.439
CLIP-only filtering	0.445	0.415
Data Engine	**0.549**	**0.517**

Table 3은 각 방법으로 선별된 데이터셋으로 학습한 다운스트림 모델의 성능 비교 결과이다. 데이터 엔진으로 획득한 데이터셋이 Box mAP 0.549, Mask mAP 0.517로 가장 높은 성능을 보였다. 한편 키워드 크롤링이 CLIP-only 필터링보다 다운스트림 성능이 다소 높게 나타났는데, 이는 키워드 크롤링이 목표 클래스와 무관한 이미지를 다수 포함하더라도 이후의 오토 레이블링 과정에서 무관한 이미지 내에 목표 객체가 검출되지 않아 훈련 레이블 생성 대상에서 1차적으로 제외되기 때문이다. 즉, 오토 레이블링 단계에서 무관한 이미지가 걸러지는 효과로 인해 실제 학습에 사용되는 레이블이 상대적으로 정제되면서 다운스트림 성능이 보존된 것으로 해석된다. 다만 사람이 전체 데이터를 검수하는 관점에서는 키워드 크롤링 방식이 훨씬 많은 양의 무관한 데이터를 포함하므로, 검수와 필터링에 소요되는 시간과 비용이 크게 증가한다.

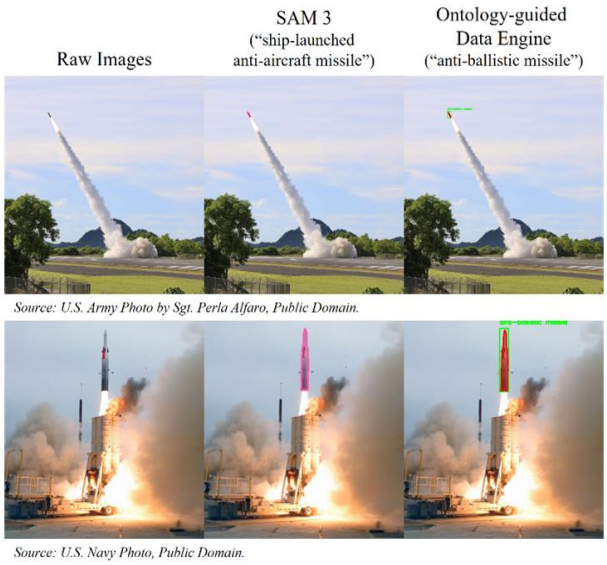


Fig. 7. Qualitative comparison results between SAM 3 and the proposed framework

4. Labeling Results

본 절에서는 제안하는 프레임워크가 텍스트만으로는 구분이 어려운 미세 객체를 얼마나 정교하게 식별하는지 검증하기 위해, 최신 비전 파운데이션 모델인 SAM 3[9]와 정성적 비교 평가를 수행하였다. 비교 대상은 '탄도탄 요격 미사일(Anti-ballistic missile)'과 '함대공 미사일(Ship-launched anti-aircraft missile)'이다. 두 객체는 모두 원통형의 동체를 가진 '미사일'이라는 공통점을 공유하여, 텍스트 기반 모델이 혼동하기 쉬운 대표적인 사례이다.

Fig. 7은 실제 'Anti-ballistic missile' 이미지를 입력으로 주었을 때, 두 모델의 레이블링 결과를 보여준다. 실험은 두 모델에게 'Ship-launched anti-aircraft missile'을 탐지하도록 유도하거나, 각 모델의 방식대로

www.kci.go.kr

최적의 클래스를 할당하게 했을 때의 차이를 분석하였다.

SAM 3는 'Ship-launched anti-aircraft missile'이라는 텍스트 프롬프트가 주어졌음에도 불구하고, 입력된 'Anti-ballistic missile'에 대해 분할 마스크를 생성하였다. 이는 모델이 'Ship-launched'이라는 수식어의 맥락(해상, 수직 발사대 등)을 인지하지 못하고, '미사일'이라는 단어가 가지는 일반적인 시각적 특징에만 과도하게 집중했기 때문이다. 결과적으로 SAM 3는 두 미사일의 미세한 차이를 구분하지 못하고 동일한 범주로 처리하는 한계를 보였다.

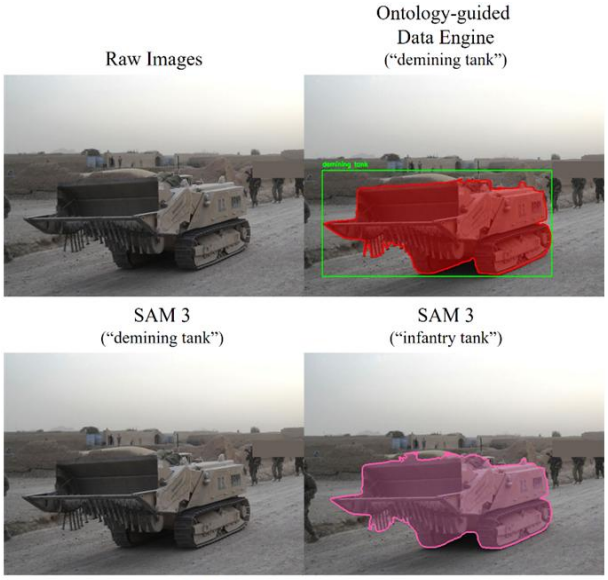


Fig. 8. Segmentation results on a rare class

반면, 본 프레임워크는 텍스트가 아닌 참조 이미지의 시각적 프로토타입을 기준으로 필터링을 수행한다. 본 프레임워크는 입력 이미지의 시각적 특징이 'Ship-launched anti-aircraft missile'의 프로토타입과 불일치함을 감지하여 이를 배제하고, 온톨로지 상에서 정확히 매칭되는 'Anti-ballistic missile' 클래스로 레이블링을 수행하였다.

이는 제안하는 방법론이 텍스트의 모호성을 넘어 시각적 특징을 기반으로 두 객체를 별개의 클래스로 명확히 분리하고, 각 클래스에 특화된 고유한 마스크를 생성할 수 있는 미세 식별 능력을 갖추었음을 입증한다.

또한, 본 연구는 텍스트 데이터가 부족한 희귀 클래스에 대한 식별 성능을 추가로 검증하였다. 실험 대상은 특수 목적 차량인 '지뢰 제거 전차(Demining tank)'와 일반적인 '보병 전차(Infantry tank)'이다.

Fig. 8은 실제 'Demining tank' 이미지를 입력으로 주었을 때의 모델별 결과를 보여준다. SAM 3는 정확한 클래스 명을 텍스트 프롬프트로 입력했음에도 불구하고, 해당 객체를 전혀 탐지하지 못했다. 이는 학습 데이터 부족으로 인해 'Demining'이라는 텍스트 개념과 시각적 특징을 연결하지 못한 결과이다. 또한, 'Infantry tank'라는 오답 프롬프트를 입력하자 해당 객체를 즉시 마스킹하는 오류가 발생했다. 이는 모델이 객체의 구체적 특징(지뢰 제거 장비 등)보다는 'Tank'라는 일반적인 형상에 편향되어 있음을 시사한다.



Fig. 9. Segmentation results using the proposed framework

반면, 본 프레임워크는 텍스트 인코더에 의존하지 않고, 위키미디어에서 확보된 'Demining tank'의 시각적 프로토타입을 매칭하여 해당 객체를 정확히 탐지하고 마스킹하였다. 또한, Fig. 9에서 볼 수 있듯, 일반적인 'Infantry tank' 이미지가 입력되었을 때도 온톨로지 상의 해당 클래스로 정확히 분류해냄으로써, 희귀 클래스와 일반 클래스 모두에 대해 강건한 레이블링 능력을 입증하였다.

추가적으로, 레이블링 품질을 정량적으로 비교하기 위해, 250장 테스트셋을 정답으로 하여 자동 생성된 바운딩 박스와 분할 마스크의 정확도를 평가하였다. 이때 SAM 3와의 비교 공정성을 확보하기 위해, 텍스트 프롬프트만 사용하는 기본 조건에 더해 본 프레임워크와 동일하게 위키미디어 참조 이미지를 활용하는 SAM 3 변형 조건을 추가

로 구축하였다. SAM 3는 예시(Exemplar) 프롬프트를 단일 이미지 내부의 박스로 전달받는 구조이므로, 외부 참조 이미지를 활용하기 위해 참조 이미지와 질의 이미지를 나란히 결합한 통합 이미지를 생성한 뒤 참조 이미지 영역을 예시 박스 프롬프트로 지정하여 모델을 사용하고, 질의 이미지 영역에서의 예측 결과만을 추출하였다.

Table 4. Auto-labeling quality (top) and per-class Box AP distribution (bottom)

Method	Box mAP@50:95	Mask mAP@50:95
SAM 3 (text-only)	0.616	0.638
SAM 3 (ref. imgs.)	0.656	0.666
Data Engine	**0.826**	**0.839**

Method	Per-class Box AP std (min-max)
SAM 3 (text-only)	0.260 (0.139-0.966)
SAM 3 (ref. imgs.)	0.270 (0.006-0.944)
Data Engine	**0.182 (0.309-1.000)**

Table 4에서 볼 수 있듯, 텍스트 프롬프트만 사용한 기본 SAM 3의 Box mAP는 0.616, Mask mAP는 0.638이었으며, 참조 이미지를 시각 프롬프트로 활용할 경우 각각 0.656과 0.666으로 일부 상승하였다. 그러나 이는 데이터 엔진의 레이블링 품질인 Box mAP 0.826, Mask mAP 0.839에는 현저히 미치지 못한다.

클래스별 Box AP의 분포 역시 비슷한 경향을 보여준다. 데이터 엔진의 클래스 간 표준편차는 0.182(최소 0.309, 최대 1.000)로, 프롬프트만 사용한 기본 SAM 3의 표준편차 0.260 및 참조 이미지를 사용한 SAM 3의 표준편차 0.270보다 작았다. 또한, 텍스트 프롬프트 SAM 3의 최저 성능 클래스는 demining tank(0.139), 참조 이미지 SAM 3의 최저 성능 클래스는 ship-launched anti-aircraft missile(0.006)로, 정성 평가에서 관찰되었던 혼동 양상이 정량적 결과에서도 동일하게 확인되었다.

이러한 결과는 범용 파운데이션 모델에 참조 이미지를 In-context 프롬프트로 입력하는 단순 결합만으로는 세부 클래스 간의 미세한 차이를 포착하기 어렵다는 점을 시사하며, 국방 도메인과 같이 미세한 클래스 분포를 갖는 영역에서는 참조 이미지를 온톨로지 기반으로 수집하고 시각 프로토타입으로 정제하는 파이프라인 구축이 필수적임을 보여준다.

한편, 데이터 엔진 프레임워크의 실패 사례도 존재하는데 이는 성능이 가장 낮게 측정된 ship-launched anti-aircraft missile(0.309) 클래스에서 관찰되었다. 미사일 발사 시 발생하는 화염과 연기에 가려진 소형 비행체를 식별하지 못하고 발사 구조물만 레이블링하는 미탐과, 복잡한 함정 구조물의 일부 영역을 미사일로 오인하는 오탐이 낮은 성능의 주요 원인으로 판단된다.

5. Computational Cost

Table 5는 파이프라인 단계별 계산 비용을 보여준다. 후보 이미지 수집은 클래스당 약 21초, 프로토타입 생성 및 의미론적 필터링은 이미지당 76밀리초, 오토 레이블링은 필터링을 통과한 이미지당 739밀리초가 소요된다. 이를 12개 클래스, 클래스당 100장 규모의 파일럿 실험에 적용하면 후보 수집 약 4분, 필터링 약 1.5분, 오토 레이블링 약 8분으로 전체 프로세스에 총 14분 정도가 소요되며, 이는 클래스당 약 70초 만에 100장 규모의 정제된 레이블링 데이터셋을 구축할 수 있음을 의미한다. 따라서 본 데이터 엔진 프레임워크는 제한된 컴퓨팅 자원을 가진 폐쇄망 환경에서도 충분한 실용성을 갖는다. 본 실험에서, 온톨로지 및 참조 이미지 수집은 위키데이터 및 위키미디어 API 응답 속도에 의존하는 1회성 사전 작업이므로 계산 비용 산정에서는 제외하였다.

Table 5. Computational cost

Stage	Processing time
Candidate image collection	21 s / class
Prototype & semantic filtering	76 ms / image
Auto-labeling	739 ms / image

V. Limitations

본 연구의 누적식 어블레이션은 키워드 크롤링, CLIP-only 필터링, 제안 방법 간의 비교를 통해 의미론적 필터링과 시각 프로토타입의 기여를 분리하였으나, 온톨로지 자체의 독립적인 기여는 분리하지 못하였다. 제안 프레임워크에서 온톨로지는 수집 대상 클래스를 계층적으로 구조화 및 정의하고, 해당 하위 개념을 따라 참조 이미지 수집을 유도하는 역할을 하며, 시각 프로토타입은 수집된 참조 이미지로부터 생성된다. 즉, 온톨로지의 결과가 후속 단계의 입력이 되는 순차 의존적 구조이므로, 온톨로지의

www.kci.go.kr

기여가 시각 프로토타입의 기여와 결합된 형태로 측정된다. 이는 본 프레임워크의 한계로, 온톨로지 확장 범위에 따른 최종 성능 변화 분석은 향후 과제로 남긴다.

VI. Conclusions

본 연구에서는 데이터 확보가 제한적이고 높은 전문성이 요구되는 국방 도메인의 특수성을 극복하기 위해, 온톨로지 기반의 데이터 엔진(Ontology-Guided Data Engine) 프레임워크를 제안하였다. 제안된 프레임워크는 위키데이터 지식 그래프를 통해 군사 객체의 개념을 정교하게 정의하고, 최신 비전 파운데이션 모델들의 제로샷 능력을 결합하여 데이터의 수집부터 정제, 레이블링에 이르는 전 과정을 자동화하였다. 12개 군사 클래스에 대한 정량 평가에서 제안 방법은 키워드 크롤링 및 CLIP-only 필터링 대비 가장 높은 데이터 선별 품질을 기록하였고, 참조 이미지를 활용한 SAM 3 대비 우수한 오토 레이블링 품질을 달성하였으며, 선별된 데이터로 학습한 다운스트림 모델에서도 가장 높은 탐지 및 분할 성능을 보여주었다. 이를 통해, 본 프레임워크는 기존 텍스트 프롬프트 기반 모델들이 구분하지 못하는 미세한 시각적 차이를 프로토타입 기반 필터링을 통해 효과적으로 식별함을 확인하였다. 향후 연구에서는 구축된 온톨로지를 제원(Specification) 및 전술적 설명(Tactical Description) 텍스트와 결합하여, 시각 정보와 언어 정보가 통합된 멀티모달 데이터셋 구축 엔진으로 확장하고자 한다.

ACKNOWLEDGEMENT

This work was supported by Korean Government.

REFERENCES

[1] J. Redmon, et al., "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016. DOI: 10.1109/CVPR.2016.91

[2] T. Y. Lin, et al., "Bilinear CNN Models for Fine-Grained Visual Recognition," IEEE International Conference on Computer Vision (ICCV), pp. 1449-1457, 2015. DOI: 10.1109/ICCV.2015.170

[3] K. He, et al., "Mask R-CNN," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 2, pp. 386-397, Feb. 2020. DOI: 10.1109/TPAMI.2018.2844175

[4] T. Y. Lin, et al., "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision (ECCV), pp. 740-755, 2014. DOI: 10.1007/978-3-319-10602-1_48

[5] J. Deng, et al., "ImageNet: A Large-Scale Hierarchical Image Database," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255, 2009. DOI: 10.1109/CVPR.2009.5206848

[6] D. Vrandecic, et al., "Wikidata: A Free Collaborative Knowledgebase," Communications of the ACM, Vol. 57, No. 10, pp. 78-85, Oct. 2014. DOI: 10.1145/2629489

[7] S. Liu, et al., "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," European Conference on Computer Vision (ECCV), pp. 38-55, 2024. DOI: 10.1007/978-3-031-72970-6_3

[8] A. Kirillov, et al., "Segment Anything," IEEE International Conference on Computer Vision (ICCV), pp. 4015-4026, 2023. DOI: 10.1109/ICCV51070.2023.00371

[9] N. Carion, et al., "SAM 3: Segment Anything with Concepts," International Conference on Learning Representations (ICLR), 2026. DOI: 10.48550/arXiv.2511.16719

[10] G. S. Xia, et al., "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3974-3983, 2018. DOI: 10.1109/CVPR.2018.00418

[11] P. Zhu, et al., "Vision Meets Drones: A Challenge," arXiv preprint arXiv:1804.07437, 2018. DOI: 10.48550/arXiv.1804.07437

[12] D. Lam, et al., "xView: Objects in Context in Overhead Imagery," arXiv preprint arXiv:1802.07856, 2018. DOI: 10.48550/arXiv.1802.07856

[13] E. Gundogdu, et al., "MARVEL: A Large-Scale Image Dataset for Maritime Vessels," Asian Conference on Computer Vision (ACCV), pp. 165-180, 2016. DOI: 10.1007/978-3-319-54193-8_11

[14] Z. Shao, et al., "SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection," IEEE Transactions on Multimedia, Vol. 20, No. 10, pp. 2593-2604, Oct. 2018. DOI: 10.1109/TMM.2018.2865686

[15] X. Chen, et al., "NEIL: Extracting Visual Knowledge from Web Data," IEEE International Conference on Computer Vision (ICCV), pp. 1409-1416, 2013. DOI: 10.1109/ICCV.2013.178

[16] S. K. Divvala, et al., "Learning Everything about Anything: Webly-Supervised Visual Concept Learning," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3270-3277, 2014. DOI: 10.1109/CVPR.2014.412

[17] X. Wang, et al., "Zero-Shot Recognition via Semantic

www.kci.go.kr

Embeddings and Knowledge Graphs," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6857-6866, 2018. DOI: 10.1109/CVPR.2018.00717

[18] A. Radford, et al., "Learning Transferable Visual Models from Natural Language Supervision," International Conference on Machine Learning (ICML), pp. 8748-8763, 2021. DOI: 10.48550/arXiv.2103.00020

[19] H. Xu, et al., "Demystifying CLIP Data," International Conference on Learning Representations (ICLR), 2024. DOI: 10.48550/arXiv.2309.16671

[20] T. Ren, et al., "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," arXiv preprint arXiv:2401.14159, 2024. DOI: 10.48550/arXiv.2401.14159

[21] O. Hartig, et al., "Executing SPARQL Queries over the Web of Linked Data," International Semantic Web Conference, pp. 293-309, 2009. DOI: 10.1007/978-3-642-04930-9_19

[22] M. Gong, et al., "A Review of Non-Maximum Suppression Algorithms for Deep Learning Target Detection," Symposium on Novel Photoelectronic Detection Technology and Applications, Vol. 11763, pp. 821-828, Mar. 2021. DOI: 10.1117/12.2586477

[23] G. Jocher, et al., "Ultralytics YOLOv8," Software available at *https://github.com/ultralytics/ultralytics*, 2023.

Authors



Seonghak Kim is a researcher with the Defense AI R&D Institute, Agency for Defense Development (ADD), Daejeon, South Korea. His current research interests focus on computer vision and model compression, with particular emphasis on knowledge distillation, class-incremental learning, small object detection, and vision foundation models.

www.kci.go.kr