캡스톤 디자인 I 최종결과 보고서

프로젝트 제목(국문): 항공영상 기반의 제로샷 객체 탐지 연구
프로젝트 제목(영문): Zero-shot Object Detection Based on Aerial Imagery

프로젝트 팀(원): 학번: 20201735 이름: 박우진
프로젝트 팀(원): 학번: 20222019 이름: 김다빈

1. 중간보고서의 검토결과 심사위원의 '수정 및 개선 의견'과 그러한 검토의견을 반영하여 개선한 부분을 명시하시오.
- 해당사항 없음

2. 기능, 성능 및 품질 요구사항을 충족하기 위해 본 개발 프로젝트에서 적용한 주요 알고리즘, 설계방법 등을 기술하시오.

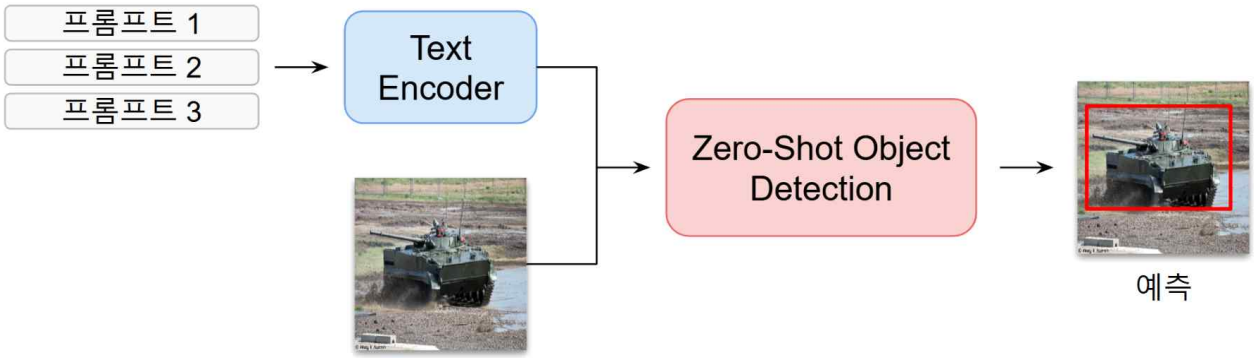
2-1) 군용 객체 이미지 제로샷 탐지를 위한 Open-Vocabulary Object Detection (OVD) 및 Vision-Language Model(VLM) 모델 기반 구조 설계
- 다양한 군용 객체(전차, 장갑차 등)에 대해 텍스트 프롬프트 기반 제로샷 탐지가 가능하도록, YOLO-World 등 Open-Vocabulary Object Detection 모델과 Qwen 등 VLM을 활용한 Zero-Shot Object Detection 구조를 설계하고 실험하였다.

2-2) 보안상 확보하기 어려운 군용 객체 드론 영상을 대체하기 위한 합성 이미지 생성
- Stable Diffusion 모델과 ControlNet을 결합한 이미지 합성 기법을 통해 대체 데이터를 생성하였다.
- 실제 데이터가 부족한 군용 객체에 대해, 다양한 배경(도로, 초원, 산 등)과 시점(드론뷰 포함)에서의 합성 이미지를 생성함으로써 데이터 다양성과 탐지 성능 향상을 도모하였다.

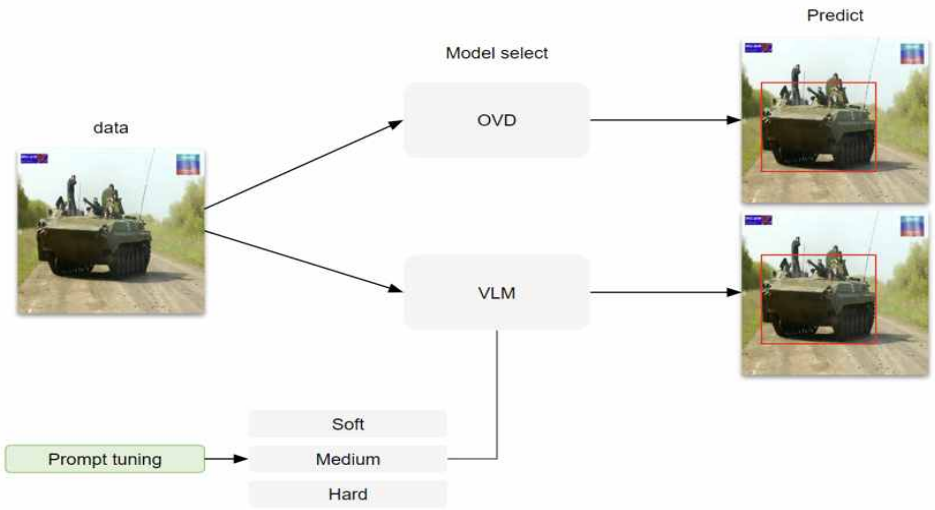
2-3) 군용 객체 제로샷 탐지 성능을 향상시키기 위한 Zero-Shot Object Detection 논문 탐색 및 재현 실험
- YOLO-World, YOLO-UniOW, Grounding DINO 등 최신 Zero-Shot Object Detection 모델을 활용하여 실험을 수행하고, 성능 재현 및 군용 객체 탐지 성능 향상에 기여할 수 있는 요소를 분석하였다.

3. 요구사항 정의서에 명세된 기능 및 품질 요구사항에 대하여 최종 완료된 결과를 기술하시오.

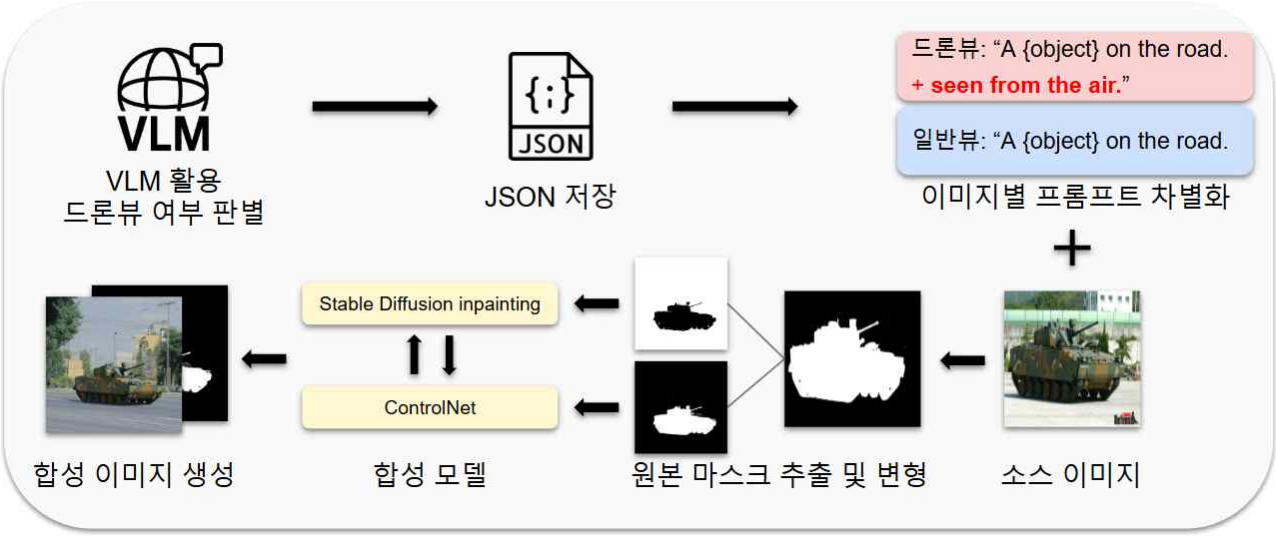
3-1) 제로샷 객체 탐지 파이프라인



3-2) 시각-언어 모델(VLM)기반 군용 객체 탐지 파이프라인

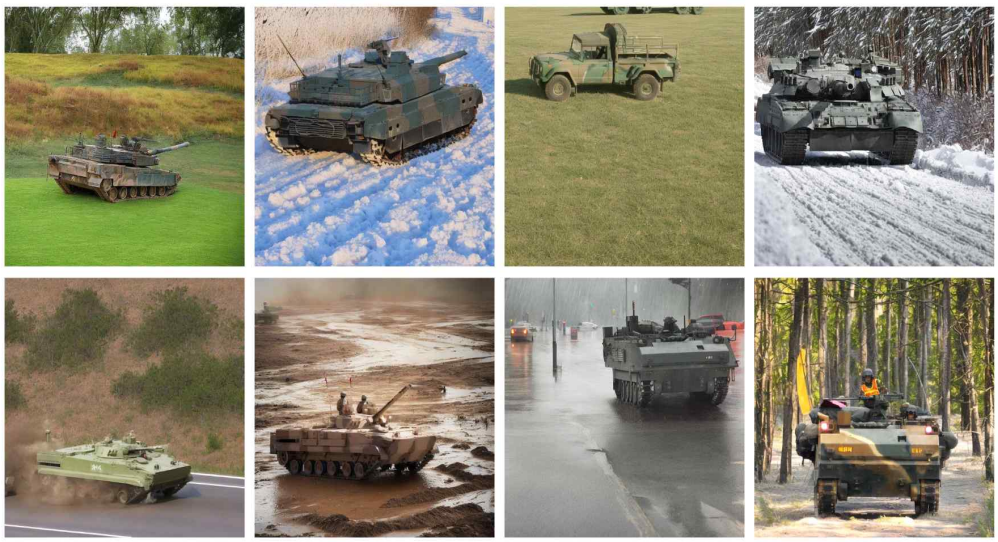


3-3) 합성 기법 파이프라인



3-4) 제안 기법의 합성 결과

- 이외 합성 이미지



3-5) 기존 객체 탐지 모델 구조 및 특징 비교 분석
- 비교 연구를 위한 최신 제로샷 탐지 모델 탐색

	논문 제목	학술지 /학술대회	발행년도	탐지 방식	특징
1	YOLO-World: Real-Time Open-Vocabulary Object Detection	CVPR	2024	Zero-shot	1. CSPDarknet 기반 경량 backbone을 사용하여 feature 추출 2. Cross-attention 모듈을 neck과 head에 삽입하여 임베딩 수준에서 클래스 융합
2	Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection	ECCV	2024	Zero-shot	1. Swin Transformer 기반 vision backbone으로 high-resolution feature 표현 강화 2. 사전 학습된 BERT/CLIP 텍스트 인코더로 자연어 프롬프트를 처리 3. 디코더에 cross-attention 레이어를 추가해 텍스트 쿼리와 비전 쿼리 융합
3	YOLO-UniOW: Efficient Universal Open-World Object Detection	arxiv	2024	Zero-shot	1. YOLO-World 를 Open-World로 확장 2. Detector를 yolov8에서 yolov10으로 변경, 효율성 개선 3. image-text fusion의 단점 개선, AdaDL을 통한 text encoder LORA fine-tuning 4. unknown 객체 탐지를 위한 wild card learning 5. YOLO-World 보다 높은 성능 달성

- Vision-Language 모델

	논문 제목	학술지 /학술대회	발행년도	특징
1	Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks	CVPR	2024	1. Image captioning, object detection, visual grounding, segmentation 등 다양한 Vision-Language 작업을 간단한 텍스트 프롬프트 하나로 수행할 수 있는 모델 2. 대규모 FLD-5B 데이터셋을 활용해 학습되었으며, 이를 통해 파인튜닝 없이도 준수한 제로샷 성능을 보여줌
2	Qwen2.5 Technical Report	-	2024	1. 광범위한 지침 학습으로 복잡하고 구체적인 작업도 프롬프트 하나만으로 정확히 수행할 수 있어, 다양한 포맷의 출력을 즉시 생성 가능 2. 별도 추가 학습 없이도 제로샷 탐지를 지원해, 입력한 프롬프트에 따라 원하는 결과 형태를 즉시 반환 가능

3-6) 군용 차량 탐지 성능 향상을 위한 프롬프트 튜닝과 합성 데이터 전략 기법 제안 및 논문 작성
[1] 프롬프트 튜닝을 통한 시각-언어 모델 기반 군용 객체 제로샷 탐지 성능 향상 연구
- 시각-언어 모델(VLM) 기반 군용 객체 탐지 성능 향상을 위한 프롬프트 튜닝 실험

등록자 정보

* 등록자 성명 김다빈

* 등록자 이메일 20222019@edu.hanbat.ac.kr

논문 정보

* 논문 제목 프롬프트 튜닝을 통한 시각-언어 모델 기반 군용 객체 제로샷 탐지 성능 향상 연구

* 초록 본 논문은 군용 객체 이미지에 대한 제로샷 객체 탐지 성능을 향상시키기 위한 기법을 제안한다. 전통적 Open-Vocabulary Detection(OVD) 모델과 Vision-Language Model(VLM)을 비교하고 특히 프롬프트 튜닝이 탐지 성능에 미치는 영향을 실험적으로 분석하였다. Roboflow Russian-military 데이터셋을 COCO 형식 GT로 재구성하여 평가한 결과 Grounding DINO가 OVD 계열 모델 중 mAP 30.4%로 가장 우수한 성능을 보였으며, VLM 계열 모델인 Qwen2.5-VL-7B에 3단계 프롬프트 튜닝을 적용한 조건에서 Soft는 mAP 33.3%, Medium은 mAP 42.7%, Hard는 mAP 46.8%로 단계적으로 성능이 향상하는 경향을 나타냈다. 이러한 결과는 VLM 기반 제로샷 탐지에서 프롬프트의 구체성과 엄격성이 검출 정확도에 결정적 요소임을 실증하며 군용 객체와 같은 특수 도메인에의 적용 가능성과 확장성을 제시한다.

2025-05-03 김다빈 학술대회 구두 프롬프트 튜닝을 통한 시각-언어 모델 기반 군용 객… 승인 논문파일 저작권파일 상세보기

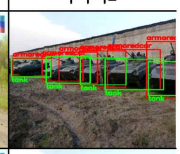
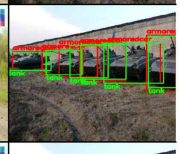
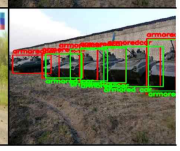
모델	tank_AP	armored car_AP	military truck_AP	mAP
YOLO-World	12.9	35.8	8.2	19.0
Grouding-DINO	23.5	40.1	27.5	30.4
Florence-2	17.6	39.7	6.3	21.2
Qwen_S	27.5	22.3	50.2	33.3
Qwen_M	35.3	42.6	50.1	42.7
Qwen_H	38.1	45.0	57.4	46.8

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

모델	이미지1	이미지2
YOLO-World		
Grounding DINO		
Florence-2		
Qwen_S		
Qwen_M		
Qwen_H		

모델	이미지1	이미지2
Qwen_S		
Qwen_M		
Qwen_H		

[2] 확산 모델 기반 배경 재구성 합성 데이터를 활용한 데이터 부족 환경에서의 군용 차량 객체 탐지 연구
- 군용 차량 데이터 부족 문제를 보완하기 위한 데이터 합성 기법 제안

등록자 정보

* 등록자 성명: 박우진

* 등록자 이메일: 20201735@edu.hanbat.ac.kr

논문 정보

* 논문 제목: 확산 모델 기반 배경 재구성 합성 데이터를 활용한 데이터 부족 환경에서의 군용 차량 객체 탐지 연구

* 초록: 본 연구는 보안상 공개가 제한된 군용 차량 이미지 데이터의 부족 문제를 해결하기 위해 Stable Diffusion 기반 Inpainting과 ControlNet 조건부 합성을 결합한 고품질 합성 데이터 생성 파이프라인을 제안한다. 객체의 형태는 그대로 유지하되 배경만 재구성하여 합성 이미지를 확보하고, 실제·합성 데이터를 단계적으로 융합하여 활용하였다. 실험 결과, YOLOv8 모델의 mAP(0.5:0.95)는 베이스라인(58.4%) 대비 합성 데이터 추가 시 60.4%로, 합성 적용 전략 최적화 시 65.6%로 7.2% 향상되었다. 이러한 성능 개선은 Diffusion 기반 합성 데이터가 군용 차량 탐지 모델의 일반화 능력을 효과적으로 강화함을 입증하며, 본 기법은 데이터 부족 및 도메인 간 격차 문제를 겪는 다양한 객체 탐지 과제에 대한 새로운 해결책을 제시한다.

2025-05-03	박우진	학술대회	구두	확산 모델 기반 배경 재구성 합성 데이터를 활용한 데…	승인	논문파일 저작권파일	상세보기

Method	Train data	Augmentation	Test mAP(0.5:0.95) (RDA대비 성능 향상률)
RDA (Real+DefaultAugmentation)	실제 데이터	Default 증대법 (YOLO기본)	58.4%
RFDA (Real+Fusion+DefaultAugmentation)	실제+ 합성 데이터	Default 증대법 (YOLO기본)	60.4% (+2.0%)
RFRA (Real+Fusion+ReducedAugmentation)	실제+ 합성 데이터	Default 증대법에서 Mosaic와 HSV 증대법만 제거	61.5% (+3.1%)
RFSA (Real+Fusion+SplitAugmentation)	실제+ 합성 데이터	실제: Default 증대법에서 Mosaic와 HSV 증대법만 제거 합성: Horizontal flip	**65.6%** **(+7.2%)**

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

Ground Truth | RDA 모델 | RFDA 모델 | RFRA 모델 | RFSA 모델

3-7) 소프트웨어중심대학 AI 경진대회 참가: 4im 팀, K2E 팀

AI부문	1	AirLab_A	50만원	디지털경진대회 참가
	2	AirLab_B	40만원	디지털경진대회 참가
	3	4im	30만원	디지털경진대회 참가
	4	K2E		디지털경진대회 참가
	5	KI Lab		디지털경진대회 참가
	6	VML		디지털경진대회 참가
	7	오솝		디지털경진대회 참가

4. 구현하지 못한 기능 요구사항이 있다면 그 이유와 해결방안을 기술하시오.

최초 요구사항	구현 여부(미구현, 수정, 삭제 등)	이유(일정부족, 프로젝트 관리미비, 팀원변동, 기술적 문제 등)
T-80, K2 등 세부 클래스명으로 탐지	상위 클래스(tank, armored car, military truck) 탐지로 대체	T-80, K2와 같은 세부 군용 객체 간에는 외형적 차이가 미세하고 촬영 조건(해상도, 각도 등)에 따라 구분이 어려워 탐지 성능이 불안정했음. 따라서 보다 안정적인 탐지를 위해 상위 클래스 단위로 통합함
OVD 모델을 활용하여 군용 객체 탐지	VLM을 활용하여 객체 탐지	OVD 모델만으로는 군용 객체의 복잡한 외형이나 맥락적 정보(예: 전차 종류 간 구분)를 효과적으로 반영하기 어려웠으며, 이를 보완하기 위해 텍스트-이미지 정렬 능력이 뛰어난 VLM 기반 구조를 도입하여 탐지 성능을 향상시킴

5. 요구사항을 충족시키지 못한 성능, 품질 요구사항이 있다면 그 이유와 해결방안을 기술하시오.

분류(성능, 속도 등) 및 최초 요구사항	충족 여부(현재 측정결과 제시)	이유(일정부족, 프로젝트 관리미비, 팀원변동, 기술적 문제 등)
		해당사항 없음

6. 최종 완성된 프로젝트 결과물(소프트웨어, 하드웨어 등)을 설치하여 사용하기 위한 사용자 매뉴얼을 작성하시오.

6-1) 시각-언어 모델(VLM)기반 군용 객체 탐지

- 하드웨어 정보 및 파이썬 버전

Python 3.10.16
Ubuntu 22.04.5 LTS
GPU: NVIDIA RTX A6000 x 2
CPU: AMD EPYC 7543 32-Core Processor
Memory: 251G

- 사용자 매뉴얼
<라이브러리 설치>
conda create -n qwen-vl python=3.10
conda activate qwen-vl

1) PyTorch + torchvision + CUDA 지원
conda install pytorch=2.6.0 torchvision=0.21.0 pytorch-cuda=12.4 -c pytorch -c nvidia

2) 그 외 필수/권장 패키지
pip install transformers==4.49.0.dev0 \
            qwen-vl-utils \
            pillow \
            tqdm \
            huggingface-hub \
            accelerate \
            sentencepiece \
            safetensors

<data.py>
tank, armored car, military truck 클래스로 매핑하여 test_converted.json파일로 변환

<vlm.py>
- Qwen2.5-VL-7B-Instruct 모델로 드론/지상 뷰 이미지를 입력받아 군용 차량(탱크, 장갑차, 군용 트럭)을 탐지하고, 탐지된 바운딩 박스를 원본 해상도로 되돌려 그린 뒤 이미지를 저장한 뒤 그 결과를 COCO 포맷의 JSON으로 출력한 뒤 최종적으로 GPU별로 나뉜 JSON들을 합쳐서 하나의 detection_results_coco_merged.json파일을 만드는 전체 파이프라인
- 터미널에 huggingface-cli login 명령어 실행 후 허깅페이스 access token 붙여넣기
- 해당 코드에서 이미지 폴더 경로, 출력 폴더 경로 입력 후 실행
- 예시) python vlm.py --image_dir /workspace/dabin/YOLO-World/data/russia/train --output_dir /workspace/dabin/YOLO-World/data/russia/qwen/hard

<convert.py>
detection_results_coco_merged.json파일을 COCO prediction 리스트 형식으로 변환하기 위한 스크립트

6-2) 데이터 합성 기법
- 하드웨어 정보 및 파이썬 버전
Python 3.8.6
Ubuntu 22.04.3 LTS
GPU: NVIDIA RTX A5000 x 2
CPU: AMD Ryzen 7 5800X 8-Core Processor
Memory: 48GiB

- 사용자 매뉴얼
합성 데이터 생성 방법

<qwen_drone_view.py>
- 해당 코드에서 이미지 폴더 경로, 출력 폴더 경로 입력 후 실행

- "--image_dir", type=str, default='이미지 폴더 경로', help="Directory containing the input images "
- "--output_dir", type=str, default='출력 결과 폴더 경로', help="Directory to save classification results"

파라미터 설명
1). **inputs: input_ids, attention_mask 등 입력 텐서들을 포함하는 딕셔너리
2). max_new_tokens: 새로 생성할 최대 토큰 수
3). num_beams: Beam Search의 beam 수
4). do_sample: 샘플링 기반 생성 여부
5). temperature: 확률 분포를 부드럽게 또는 날카롭게 조절. 높을수록 랜덤성 증가
6). top_p: Nucleus Sampling에서 누적 확률이 top_p 이하가 되도록 상위 토큰만 고려
7). top_k: 상위 k개의 확률 높은 토큰 중에서 샘플링
8). early_stopping: beam search 시 조건을 만족하면 조기 종료

<image_generation.py>
- 해당 코드에서 드론뷰 여부 포함 .json 경로, 소스 이미지 경로, 출력 폴더 경로 설정 후 실행
- source_folder = '사용할 소스 이미지 폴더 경로'
- base_save_root = '저장될 합성 이미지 폴더 경로'

파라미터 설명
1). prompt: 생성할 이미지에 대한 텍스트 설명
2). image: 객체가 지워진(inpaint용) 원본 이미지
3). mask_image: 이미지에서 어디를 새로 생성할지 나타내는 마스크
4). control_image: ControlNet의 입력으로 사용되는 이미지
5). num_images_per_prompt: 하나의 프롬프트로 생성할 이미지 개수
6). generator: 시드를 고정하기 위한 PyTorch 랜덤 제너레이터
7). num_inference_steps: 디퓨전 과정의 스텝 수
8). guess_mode: ControlNet의 guess mode를 사용할지 여부
9). controlnet_conditioning_scale: ControlNet의 객체 제어 강도
10). scale: 합성 객체 무작위 스케일 설정

6-3) OVD 모델 학습

<YOLO-World>
- 학습, 검증, 테스트에 사용할 images, annotations.json 준비 후 config.py 내 train_dataloader, val_dataloader, test_dataloader 수정 후 실행
- train & evaluation: /tools/dist_train.sh configs/pretrain/custom config.py 2 --amp
- inference: /tools/dist_test.sh configs/pretrain/custom config.py 2 —amp

파라미터 설명
1). num_classes: 검증 클래스 수
2). num_training_classes: 학습 클래스 수
3). max_epochs: 최대 학습 수
4). base_lr: 기본 학습률
5). load_from: 사전학습 모델 가중치 파일 경로

7. 캡스톤디자인 결과의 활용방안

본 연구는 Open-Vocabulary Object Detection(OVD) 기법을 적용하여 새로운 유형의 군용 차량이 등장해도 모델을 재훈련하지 않고도 실시간으로 탐지할 수 있도록 설계할 것이다. 이를 통해 적군 전력 변화에 즉각 대응하고 전장 정보 수집 능력을 극대화할 것이다.

군사 분야에서는 대규모 데이터 확보가 어려워 라벨링 비용과 시간이 많이 소요된다. 본 연구는 이미지 합성 기법을 활용하여 다양한 조명·기상·촬영 각도 환경을 반영한 군용 객체 데이터를 생성함으로써 데이터 부족 문제를 해결하고 모델의 일반화 성능을 향상시킬 것이다. 합성 데이터와 실제 드론 영상을 혼합 학습하여 모델이 다양한 상황에서도 안정적으로 탐지할 수 있게 할 것이다.

군사 분야 활용 사례로는 드론 기반 실시간 전장 감시 및 정찰을 통해 적군 차량 배치를 분석하고 전략적 의사결정에 활용하는 것이 있다. 또한 새로운 차량이 도입될 때에도 재훈련 없이 즉시 탐지하여 정보 우위를 확보할 수 있을 것이다. 무인 감시 시스템과 결합하여 국경 지역 침입 감시에도 활용할 수 있을 것이다.

민간·공공 분야에서는 재난 대응 시 피해 지역의 차량 및 중요 인프라를 탐지하고 도심 교통 흐름 분석 및 사고 대응에 활용할 수 있을 것이다. 보안 감시 카메라 시스템에 적용하여 위험 탐지 기능을 강화하고, 자율주행 차량이나 무인 항공기(UAV)에 탑재하여 객체 탐지·인지 기능을 강화할 것이다.

본 연구 결과로 개발된 OVD 기반 드론 영상 객체 탐지 알고리즘, 합성 데이터 생성 기법, 경량화 모델 설계 등은 특허 및 실용신안 출원을 검토할 가치가 있을 것이다. 국방·방산 기업과 협업하여 군사 장비 개발 및 방산 시스템에 적용하고, 공공기관·민간 기업과 협력하여 스마트 감시 시스템과 자율주행 솔루션에 상용화할 수 있을 것이다.