대한조선학회논문집 Research Paper
Journal of the Society of Naval Architects of Korea
eISSN:2287-7355, Vol. 63, No. 2, pp. 101-114, April 2026
https://doi.org/10.3744/SNAK.2026.63.2.101

시각-언어 모델을 활용한 자율운항선박의 RoI 기반 데이터 효율적 항로표지 분류 기법 연구

임선혁[1] · 서진혁[1] · 김시원[1] · 정성현[1] · 김연수[1] · 조현재[2] · 박종용[1,†]
국립부경대학교 마린융합디자인공학과 조선해양공학전공[1]
한국해양과학기술원 부설 선박해양플랜트연구소[2]

A Study on RoI-based Data-efficient AtoN Classification for MASS using VLM

Sun-Hyuck Im[1] · Si-Won Kim[1] · Seong-Hyeon Jung[1] · Jin-Hyeok Seo[1] · Yeon-Soo Kim[1] · Hyun-Jae Jo[2] · Jong-Yong Park[1,†]
Department of Marine Design Convergence Engineering, Pukyong National University[1]
Korea Research Institute of Ships & Ocean Engineering[2]

This study proposes a RoI-based, data-efficient fine-grained Aids to Navigation (AtoN) classification method using vision-language models (VLMs) for Maritime Autonomous Surface Ship (MASS). The reliability of the Electronic Chart Display and Information System (ECDIS) can be limited by operating anomalies and discrepancies between charted and actual environments, motivating camera-based situational awareness to support human watch-keeping. AtoNs, which are crucial indicators for coastal navigation, are typically observed as a distant and small-scale objects, making large-scale labeled data collection difficult and degrading full-frame classification due to background dominance. To address this, we focus on RoI-based classification under limited supervision and compare a supervised YOLOv12 classifier baseline with CLIP (Contrastive Language-Image Pre-training). CLIP maximizes data efficiency through domain-specific prompt engineering grounded in IALA Region B attributes and LoRA-based few-shot tuning. Experiments on Virtual RobotX (VRX) simulation datasets under clear and foggy conditions and on real-sea RoI images demonstrate that the proposed VLM-based classifier achieves robust performance with limited training samples and maintains higher robustness under degraded visibility. These results suggest an effective direction for practical, data-efficient AtoN classification in maritime environments via RoI-based preprocessing and parameter-efficient VLM adaptation.

Keywords : MASS(Maritime Autonomous Surface Ships, 자율운항선박), AtoN(Aids to Navigation, 항로표지), YOLO(You Only Look Once, 딥러닝 모델), VLM(Vision-Language Model, 시각-언어 모델), ROS(Robot Operating System, 로봇 운영 체제)

1. 서 론

자율운항선박(MASS)의 운용 환경은 대양을 넘어 복잡한 연안 및 항만으로 확대되고 있다. 주변 환경에 대한 정확한 상황 인식은 안전 항해의 성패를 가르는 기초적이며 핵심적인 요소이다. 그러나 현재 선박 항해의 주력 시스템인 ECDIS(Electronic Chart Display and Information System)는 실제 해상 환경과의 괴리, 정보 갱신 지연과 같은 시스템 자체의 오류 가능성으로 인해 완전한 신뢰를 담보하기 어렵다. 국제해사기구(IMO)는 ECDIS의 운용상 안전성을 확보하기 위하여 초기 성능 기준(Resolution A.817(19), 1995)을 시작으로 지속적인 개정을 수행해 왔으나, 2022년에 승인된 MSC.1/Circ.1503/Rev.2에서는 ECDIS가 복잡한 항로표지(AtoN: Aids to Navigation)나 수중 위험물을 올바르게 표시하지 못하는 운용 이상 현상이 상존함을 경고하고 있다(IMO, 2022).

이러한 한계는 국제해상충돌예방규칙(COLREGs)에서 요구하는 견시의 중요성을 다시금 환기시킨다. COLREGs의 규정(Part B, Rule 5)은 모든 선박이 당시 상황에 적합한 모든 이용 가능한 수단을 활용하여 주변 상황과 충돌 위험을 평가할 것을 요구하며,

Received : 30 January 2026 | Revised : 23 February 2026 | Accepted : 4 March 2026
† Corresponding author : Jong-Yong Park, jongyong.park@pknu.ac.kr

Rule 4에 따라 이러한 경계 의무는 모든 시계 상태에 적용된다(IMO, 1972). MASS가 법적 요구사항을 충족하고 실환경에서 안전하게 운용되기 위해서는, ECDIS의 심볼 정보에만 의존하지 않고 항해사의 견시를 지원할 수 있는 시각 센서를 활용하여 실제 환경을 독립적으로 검증하는 과정이 필수적으로 요구된다. 이러한 요구에 부응하여, 최근에는 카메라 영상 데이터와 딥러닝 기술을 활용하여 타 선박을 탐지하고 식별하려는 연구가 활발히 진행되고 있다.

카메라 기반 선박 인지 연구는 주로 충돌 회피와 해상교통 감시를 위한 선박 검출·추적 성능 향상에 초점을 맞추어 발전해 왔다(Perera, 2019; Nam et al., 2021; Kaur et al., 2022; Wang et al., 2024; Park et al., 2024; Yeo et al., 2025). 이러한 연구들은 공통적으로 선박 선체를 중심으로 한 객체 검출 및 분류에 집중하고 있으며, AtoN이나 수중 구조물 등 항행 환경 요소는 대개 단순 장애물 또는 배경 객체로 취급되는 경향이 있다.

하지만 해상에서의 위험 요인은 타 선박과의 충돌 외에도 경로 이탈, 저수심 구간 진입, 금지 구역 진입 등 다양한 유형으로 존재한다. 안전 항해를 위해서는 선박뿐만 아니라 AtoN, 수중 장애물 등을 식별하여 '상황의 완전한 평가'가 선행되어야 한다. 특히 연안과 항만에서는 국제항로표지협회(IALA) 규격에 따라 색상, 형상, 두표(Top mark)의 조합으로 항행 정보를 전달하는 AtoN이 실질적인 안전의 기준이 된다. 그러나 실제 AtoN은 노후, 오염, 파손, 조류에 의한 위치 이탈 등으로 인해 ECDIS 상의 정적 정보와 시각적으로 불일치할 가능성이 크며, 이는 MASS 운항 시 항로 오인이나 위험 수역 접근과 같은 치명적 결과로 이어질 수 있다.

이러한 배경에서, AtoN을 인식 및 분류하려는 연구도 일부 수행되고 있다. Pan et al. (2020)은 딥러닝 기반의 AtoN 시각 인식 모델을 통해 다수의 AtoN 유형을 고정밀로 분류하는 미세 분류 기법을 제안하였다. 야간 환경에서는 부표의 등화 특성을 자동 식별하는 방법을 연구한 사례가 있다(Han et al., 2021; Schöller et al., 2022). 또한, 레이더 신호와 항로표지 데이터를 연관시켜 항로표지 및 다중 대상을 추적하는 연구(Choi et al., 2023), 내륙수로 환경에서 AtoN을 고정밀로 검출하기 위해 You Look Only Once(YOLO)-v4를 개선한 전용 검출 시스템(Zhen et al., 2023), 스마트 AtoN 시스템에 딥러닝 기반 객체 검출을 접목하려는 시도(Sim and Chae, 2025)가 이루어지고 있다. AtoN 분류 연구는 대체로 수천 장 이상의 라벨 데이터를 필요로 하는 지도 학습에 기반하고 있어, 실제 연안·항만에서 AtoN 유형별 충분한 학습 데이터를 확보하기 어렵다는 한계가 있다. 이러한 제약을 보완하기 위한 대안으로 VLM (Vision-Language Model)을 활용할 수 있다. 특히 CLIP(Contrastive Language-Image Pre-training)은 대조형 VLM의 한 종류로 이미지와 텍스트를 공통 임베딩 공간에 정렬시키는 사전학습을 통해, 텍스트 프롬프트 기반으로 클래스의 시각적 속성을 언어적으로 명시함으로써 제로샷(Zero-shot) 및 퓨샷(Few-shot) 환경에서 데이터 효율적이다. 다만 AtoN은 원거리·소형 객체로 관측되는 경우가 많아, 전체 프레임에 대한 전역 임베딩 기반 분류에서는 배경 정보의 비중이 커져, 규정 기반의 미세 특징이 희석될 수 있다. 따라서 본 연구에서는 AtoN이 포함된 관심 영역(RoI)을 기반으로 분류 입력을 구성하여 객체 단서의 비중을 높이고, 연산 효율 측면에서도 실시간 운용 가능성을 확보하고자 한다.

본 논문의 기여는 다음과 같이 세 가지로 요약된다. (1) 기존 CNN 기반 YOLO 시리즈의 지역적 특징 추출 한계를 극복하고 원거리 소형 객체의 전역적 문맥 파악에 유리한 Attention-centric 구조의 YOLOv12를 RoI 제안 모듈 및 베이스라인으로 채택하였다. (2) CLIP의 비전 인코더로 ViT를 사용하고, AtoN의 규정 기반

Table 1 Definition of AtoN classes based on IALA Region B

Class	Body color	Top mark Shape	Meaning
CE (East cardinal mark)	Black with yellow band	Two cones (Base to base)	Navigable water to the east
CN (North cardinal mark)	Black above yellow	Two cones (Points upward)	Navigable water to the north
CS (South cardinal mark)	Yellow above black	Two cones (Points downward)	Navigable water to the south
CW (West cardinal mark)	Yellow with black band	Two cones (Point to point)	Navigable water to the west
I (Isolated danger mark)	Black with red band	Two spheres	Isolated danger
ND (New danger mark)	Blue and yellow vertical stripes	'+' shape	Newly discovered danger
P (Port mark)	Green	Single cylinder	Port side of channel
PP (Preferred channel to port)	Red with green band	Single cone	Preferred channel is to port
PS (Preferred channel to starboard)	Green with red band	Single cylinder	Preferred channel is to starboard
SW (Safe water mark)	Red and white vertical stripes	Single sphere	Navigable water all around
SP (Special mark)	Yellow	Single 'X' shape	Special area or feature
S (Starboard mark)	Red	Single cone	Starboard side of channel

속성을 텍스트로 명시하는 프롬프트 엔지니어링을 통해 제로샷 분류 성능을 향상시키는 방법론을 제시하였다. (3) 소량 데이터만으로 적용 가능한 LoRA 튜닝을 통해 퓨샷 환경에서의 성능 개선 효과를 검증하고, 맑은 날/해무 환경의 VRX 가상 데이터 및 실제 해상 데이터로 환경 변화에 대한 강건성을 평가하였다. 또한 ROS2 환경에서 처리 지연을 계측하여 CLIP 분류기의 실시간 운용 관점의 적용 가능성을 함께 논의하였다.

2. 데이터셋 구축 및 실험 구성

2.1 IALA-B 기반 AtoN 클래스 정의

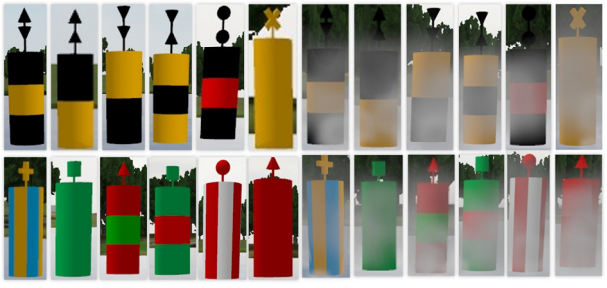
IALA의 해상부표식은 지리적 위치에 따라 전 세계를 IALA Region A와 Region B로 구분하여 적용하고 있으며, 대한민국을 포함한 북미 및 일부 아시아 지역은 IALA Region B를 채택하고 있다. IALA Region B의 핵심적인 특징은 측방표지의 색상 체계로, 항구 입항을 기준으로 우현표지는 적색, 좌현표지는 녹색으로 표시된다.

본 연구에서는 이러한 IALA-B 규정에 의거하여 데이터셋의 클래스를 정의하였다. 각 클래스는 색상, 형상, 두표의 조합에 따라 구분되며, 세부 속성 및 분류 기준은 Table 1에 정리하였다. 총 12개의 AtoN 클래스를 대상으로 분류 실험을 수행하였으며, 측방·방위·특수·독립위험표지 등 IALA-B 주요 표지군을 포함한다.

2.2 VRX 시뮬레이션 환경 및 가상 데이터 생성

본 연구에서는 VRX(Virtual RobotX)가 제공하는 Gazebo 기반 오픈소스 3D 로봇 시뮬레이션 플랫폼을 활용하였다. 해당 플랫폼은 파도, 바람, 조류 등 해상 물리 환경 모델링을 지원하며, 본 연구에서는 시각 인지 성능에 직접적인 영향을 미치는 관측 조건 변화와 해무 환경을 중심으로 가상 이미지 데이터를 생성하였다. 또한, 데이터셋의 효율적인 가공과 제안 시스템의 실시간 연동을 구현하기 위해 Ubuntu 22.04 운영체제와 ROS2 Humble 미들웨어를 기반으로 VRX 환경을 구축하였다.

Fig. 1은 Gazebo 시뮬레이터 환경에서 3D 모델로 직접 구현한 가상 AtoN의 예시를 보여준다. 해무 모델은 Gazebo의 Particle emitter를 기반으로 구현되었으며, 이는 복잡한 광학 산란 특성을 구현하기보다 다수의 미세 입자를 동적으로 생성함으로써 객체와 카메라 사이의 가시성을 물리적으로 차폐하는 방식을 취한다. 비록 실제 해상의 복잡한 산란 현상을 단순화한 모델이지만, AtoN 식별의 본질적 방해 요인인 색상 대비 감소 및 경계·특징 소실을 효과적으로 모사할 수 있다. 한편, 우천 상황의 경우 빗방울에 의한 렌즈 표면의 불규칙한 굴절 및 수적 현상(Water droplets) 등 분류 알고리즘 외적인 센서 하드웨어 변수의 영향이 지배적이므로, 본 연구에서는 인지 알고리즘 고유의 강건성 평가에 집중하기 위하여 우천 상황을 연구 범위에서 제외하였다.



(a) (b)

Fig. 1 Synthetic AtoN images generated in Gazebo, clear condition(a), foggy condition(b)

데이터의 다양성과 분류 모델의 일반화 성능을 확보하기 위하여 단일 시점에서 촬영된 이미지에 의존하지 않고, AtoN과의 상대 거리, 관측 방향, 관측 각도 및 기울기 등 관측 조건을 변화시키며 이미지 데이터를 수집하였다. WAM-V 플랫폼에 탑재된 전방 카메라 센서를 통해 시뮬레이터 내 해무 농도를 조절하여 맑은 날과 해무 조건의 AtoN을 촬영하였다. 또한 데이터셋은 전체 프레임이 아닌 AtoN RoI 기반 이미지 패치로 구성된다. 즉, 카메라로 획득된 원본 이미지에 대해 단일 클래스(AtoN)로 학습된 YOLOv12 기반 AtoN 검출기를 적용하여 출력 바운딩박스를 기반으로 생성한 RoI 패치를 분류 모델의 학습·평가 입력으로 사용하였다. 본 연구는 동일한 RoI 패치를 통한 분류 모델의 성능 비교에 초점을 둔다. 시뮬레이션에 사용된 WAM-V 및 카메라 센서의 주요 제원은 Table 2와 같으며, 전방 카메라가 탑재된 WAM-V의 형상은 Fig. 2와 같다.

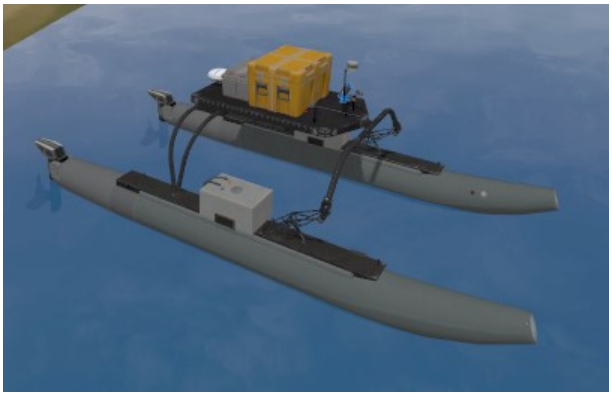


Fig. 2 WAM-V model

Table 2 Principal dimensions of WAM-V and camera specifications

Category	Parameter	Value	Unit
WAM-V	Hull length	4.9	m
	Breadth	2.5	m
	Draft	0.3	m
	Mass	151	kg
Camera	Resolution	1280×720	pixels
	Horizontal field of view	80	deg
	Frame rate	30	Hz

2.3 데이터셋 분할 및 실험 구성

본 연구에서 사용한 AtoN 이미지 데이터셋은 (1) Gazebo 기반 시뮬레이션 환경에서 생성한 가상 이미지 데이터와 (2) 실해역에서 촬영된 공개 이미지로 구성된 테스트셋으로 구성된다. 시뮬레이션-실해역 간 일반화 성능을 평가하기 위하여, 가상 데이터는 주로 모델 학습 및 검증에 활용하고, 실해역 데이터는 학습에 포함되지 않은 독립적인 테스트셋으로 이루어진다. 실해역 테스트 데이터는 클래스당 샘플 수가 제한적이나, 실제 해상의 복잡한 산란, 객체 노후화에 대한 도메인 일반화 성능을 독립적으로 평가하기 위한 목적으로 두었다.

가상 데이터셋은 클래스 간 데이터 불균형을 해소한 후 무작위 추출을 통해 학습 및 검증 세트로 분할하였다. 이때 학습 데이터는 맑은 날 조건의 가상 데이터의 RoI 패치 이미지로 구성하였으며, k-shot(k=10, 30) 설정에 따라 클래스당 각각 10장 또는 30장의 소량 학습 샘플을 사용하였다. 여기서 k=30 설정은 k=10 데이터에 데이터 증강을 적용한 새로운 패치 20장을 추가하여, 클래스당 총 30장으로 구성한 학습 데이터를 의미한다. 분류 성능 평가는 (1) 맑은 날 조건 가상 검증 데이터(클래스당 100장, 총 1200장), (2) 해무 조건 가상 검증 데이터(클래스당 100장, 총 1200장), (3) 실해역 테스트 데이터(클래스당 20장)를 순차적으로 적용하여 수행하였다.

분류 모델 간 동일한 학습 데이터셋을 공유하며, 5회의 동일한 랜덤 시드(Random seed)를 적용하여, YOLOv12 분류기와 CLIP 기반 분류 모델이 검증 이미지를 공유하도록 구성하였다. 모든 실험은 Table 3과 같은 단일 하드웨어 환경에서 수행되었다.

Table 3 Computing environment

Category	Specification
OS	Ubuntu 22.04 LTS
CPU	12th Gen Intel Core i7-12700F
GPU	NVIDIA GeForce RTX 3060 (12GB VRAM)
RAM	48GB

Table 4 Training parameters for YOLOv12 baseline

Category	Parameter	Value
Hyper-parameter	Epochs	300
	Batch	16
	Device	GPU
	Weight	YOLO12m-cls

3. RoI 기반 AtoN 분류 방법

3.1 순수 시각 기반의 AtoN 분류

본 연구에서는 CLIP 기반 분류 기법의 데이터 효율성과 환경 강건성을 검증하기 위한 비교 기준선으로, 순수 시각 기반 지도 학습 분류기를 구성하였다. 분류 단계의 베이스라인으로 YOLO 계열의 최신 아키텍처인 YOLOv12를 채택하였다.

기존의 YOLOv8부터 v11에 이르는 모델들은 주로 CNN (Convolutional Neural Network) 모듈을 기반으로 속도와 정확도의 균형을 발전시켜 왔으며, 일부 시리즈(v10, v11)에서 Attention 모듈을 도입하였으나 이는 대개 CNN 프레임워크 내에 보조적으로 삽입된 하이브리드 형태에 머물렀다. 반면, 최근 공개된 YOLOv12는 Attention-centric 구조를 전면에 내세우며 아키텍처의 진화를 이루었다(Tian et al., 2025). 구체적으로는 계산 복잡도를 낮추면서 큰 수용영역을 유지하는 A2(Area Atten- tion), Attention 도입에 따른 최적화 난이도를 완화하기 위한 R-ELAN (Residual Efficient Layer Aggregation Networks) 구조를 도입하는 등 다양한 아키텍처 개선을 통해 속도와 성능을 동시에 확보하였다. 이러한 구조적 특성은 본 연구가 제안하는 CLIP 모델 역시 Attention 메커니즘을 핵심으로 한다는 점에서, Attention 기반 기술 간의 비교를 통해 데이터 효율성 평가에 대한 대조군으로서의 설득력을 강화한다.

YOLOv12 분류기는 backbone 설계를 기반으로 분류 헤드로 치환한 분류 모델을 사용하였으며, 지도 학습 모델이 취할 수 있는 최선의 학습 조건을 제공하기 위해 사전 학습 가중치로 YOLO12m-cls를 적용하여 전이 학습을 수행하였다. AtoN 데이터셋의 RoI 패치 이미지 x_i를 입력으로 하여 YOLOv12의 Feature extractor를 기반으로 학습·평가를 수행하였다. 이 과정에서 데이터 증강(Augmentation)을 적용하여 소량 데이터 환경에서의 학습 효율을 극대화하고자 하였다. 학습 데이터셋 구축 과정에서 RoI 패치에 대한 클래스 라벨은 Roboflow 기반의 라벨링 도구를 통해 정리하였으며, 학습 내용과 학습 과정에서 적용한 세부 학습 파라미터는 Table 4와 같다.

입력 패치 x_i에 대해 YOLOv12 분류기는 C개의 AtoN 클래스에 대한 로짓(logit) $\{z_c\}_{c=1}^{C}$를 출력하며, 로짓으로부터 클래스 확률은 소프트맥스(softmax) 함수를 통해 계산된다. 입력 x_i가 클래스 c에 속할 확률 $P(y=c|x_i)$는 식 (1)과 같이 정의된다. 여기서 z_c는 클래스 c에 대한 로짓, C는 전체 AtoN 클래스 수를 의미한다.

$$P(y=c|x_i)=\frac{\exp(z_c)}{\sum_{j=1}^{C}\exp(z_j)} \quad (1)$$

최종 예측 클래스 $\hat{c}$는 식 (2)를 통해 결정한다. 학습 단계에서는 정답 클래스의 확률을 최대화하도록 교차 엔트로피 손실을 최소화하며, 클래스별 분류 손실 L_{cls}는 식 (3)과 같이 정의된다. 여기서 y_i는 해당 입력의 정답 레이블, $I(y_i=c)$는 $y_i=c$일 때 1, 그렇지 않을 때 0의 값을 갖는 지시함수이다.

$$\hat{c}=\arg\max_c P(y=c|x_i) \quad (2)$$

$$L_{cls} = -\sum_{c=1}^{C} I(y_i = c)\log P(y = c \mid x_i) \quad (3)$$

3.2 Contrastive VLM 기반의 AtoN 분류

본 연구에서 다루는 AtoN은 단순한 시각적 물체가 아니라 IALA 규정에 의해 색·형상·topmark·방위 의미의 조합으로 정의되는 의미 단위이다. 순수 지도학습 기반 분류기는 도메인 커버리지(domain coverage)를 넓히기 위해 반복적인 재학습이 요구되는 반면, CLIP은 대규모 사전학습 표현을 기반으로 zero-shot 성능을 확보할 수 있으며, LoRA와 같은 파라미터 효율적 미세 조정을 통해 소량 데이터로도 빠른 도메인 적응이 가능하다. 더불어 텍스트 프롬프트를 통해 클래스 정의를 확장할 수 있어, 지역별 표지 변형 또는 클래스 추가에 대한 유지보수 비용을 낮춘다는 점에서 운용 측면의 이점을 갖는다. 이에 본 연구는 이미지와 텍스트를 공통 임베딩 공간으로 투영하여 유사도를 계산하는 대조학습(Contrastive learning) 기반의 CLIP(Radford et al., 2021)을 AtoN 분류 모델로 채택하였다.

CLIP 알고리즘은 이미지 인코더(E_I)와 텍스트 인코더(E_T)를 통해 각각의 임베딩 벡터를 추출하고, 두 임베딩의 코사인 유사도를 최대화하도록 학습된다. 이미지 패치 x_i와 클래스 설명 텍스트 t_c에 대한 유사도는 식 (4)와 같이 정의된다.

$$S(x_i, t_c) = \frac{E_I(x_i) \cdot E_T(t_c)}{\| E_I(x_i) \| \, \| E_T(t_c) \|} \quad (4)$$

분류 단계에서는 클래스별로 특화된 프롬프트 집합 $\boldsymbol{T_c}$를 정의하고, 이미지 패치에 대해 유사도가 가장 큰 클래스를 예측한다. 각 클래스에 대해 대표 프롬프트 $t_c \in \boldsymbol{T_c}$를 두면 예측 클래스 $\hat{c}$는 식 (5)와 같이 결정된다.

$$\hat{c} = \arg\max_c S(x_i, t_c) \quad (5)$$

이미지 인코더로는 LAION-2B(Beaumont, 2022) 기반으로 대규모 사전 학습된 ViT-L/14 계열을 사용하였다. ViT는 입력 영상을 패치 단위로 임베딩한 뒤, 다층 트랜스포머 인코더를 통해 전역 문맥을 통합하며, Self-attention을 통해 이미지 내 패치 간 연관성을 학습한다. 입력 토큰은 패치 임베딩과 위치 임베딩의 합으로 구성되며, 식 (6)과 같다.

$$\boldsymbol{z}_0 = \left[x_p^1\boldsymbol{E}; x_p^2\boldsymbol{E}; \ldots; x_p^N\boldsymbol{E}\right] + \boldsymbol{E}_{pos} \quad (6)$$

해양 환경에서 해무·역광 등은 배경 노이즈를 증가시키고 객체 경계를 불명확하게 만든다. ViT 기반 인코더는 이러한 텍스처 변화에 덜 의존하고 전역적 구조·형상 단서를 통합할 수 있어, AtoN과 같이 형상 및 규정 기반 속성이 중요한 대상에 유리하게 작동한다는 장점을 가진다.

3.2.1 Zero-shot 프롬프트 엔지니어링

Zero-shot 분류 성능을 최적화하기 위해 IALA Region B 규정에 근거하여 AtoN 클래스별 색상 배치, 기하학적 형상 및 두표(top mark)를 물리적으로 묘사한 도메인 특화 앙상블(Ensemble) 프롬프트를 설계하였다. 특히 ViT-L/14 기반 이미지 인코더의 전역 문맥 파악 능력을 최대화하기 위해, 각 클래스 c에 대해 5개($K=5$)의 프롬프트 문장$\{T_{c,1}, T_{c,2}, ..., T_{c,5}\}$을 고정하여 구성하였다. 각 프롬프트는 어휘 다양성을 포함하여 몸체 패턴(예시: “Vertical stripes”, “Horizontal stripes”, “Solid color” 등), 색상 조합(예시: “Red-Green-Red”, “Solid red”, “Solid green” 등), 두표의 형상 및 방향(예시: “Two black cones pointing down”, “Green cylinder” 등) 중 일부 또는 복수 속성을 포함하도록 설계하여, 프롬프트 집합 전체가 클래스의 규정 기반 시각 단서를 포괄하도록 하였다.

클래스의 대표 텍스트 임베딩 $\bar{t}_c$는 각 프롬프트의 텍스트 임베딩을 정규화한 뒤 평균화하고, 다시 정규화하여 얻으며, 그 과정은 식 (7)과 같다. 여기서 $T_{c,k}$는 클래스 c의 k번째 프롬프트 문장을 의미한다.

$$\bar{t}_c = \frac{\frac{1}{k}\sum_{k=1}^{K}\left(\frac{E_T T_{c,k}}{\| E_T T_{c,k} \|}\right)}{\left\| \frac{1}{K}\sum_{k=1}^{K}\left(\frac{E_T T_{c,k}}{\| E_T T_{c,k} \|}\right) \right\|} \quad (7)$$

3.2.2 10-shot 튜닝

사전 학습 지식을 보존하며 IALA 규정 기반의 미세 속성을 효율적으로 주입하기 위해 모델의 전체 파라미터를 갱신하는 full fine-tuning 대신 LoRA 기반 10-shot 튜닝을 수행하였다. LoRA는 사전학습된 가중치를 고정한 채 저차원 행렬을 통해 가중치 증분만 학습함으로써, 소량 데이터 환경에서 도메인 적응이 가능하며 특정 데이터에 대한 과적합 위험을 구조적으로 낮출 수 있다.

본 연구에서는 ViT-L/14 기반 CLIP의 비전·텍스트 인코더 내부 Transformer 블록의 MLP(Multi-Layer Perceptron) 선형층에서 입력을 $\boldsymbol{x} \in \mathbb{R}^{d_{in}}$, 사전학습 가중치를 $\boldsymbol{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA 행렬을 $\boldsymbol{A} \in \mathbb{R}^{d_{in} \times r}$, $\boldsymbol{B} \in \mathbb{R}^{r \times d_{out}}$라 할 때, LoRA가 적용된 출력은 식 (8)과 같이 표현된다. 여기서 r은 LoRA rank, α는 스케일 계수, b_{fc}는 사전 학습된 선형층 bias를 의미한다.

$$f_{LoRA}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}^{\top} + (\boldsymbol{x}\boldsymbol{A}\boldsymbol{B})(\alpha/r) + b_{fc} \quad (8)$$

학습은 “정답 클래스 텍스트를 당기고, 혼동되는 타 클래스 텍스트를 밀어내는 대조학습”으로 구성하였다. 이미지 패치의 정규화된 임베딩이 식 (9)와 같을 때, 스케일링된 로짓 $z_{i,c}$는 식 (10)으로 정의된다. 여기서 γ는 LoRA 튜닝 결과에 의해 결정되는 로짓 스케일 파라미터이며, $\exp(\gamma)$는 유사도 분포의 스케일을 조절한다. 이때

Table 5 Hyper-parameter for LoRA tuning

Parameter	Value
Batch	16
Epochs	300 (Early stopping enabled)
Learning rate	1e-5
LoRA rank	2
α	4

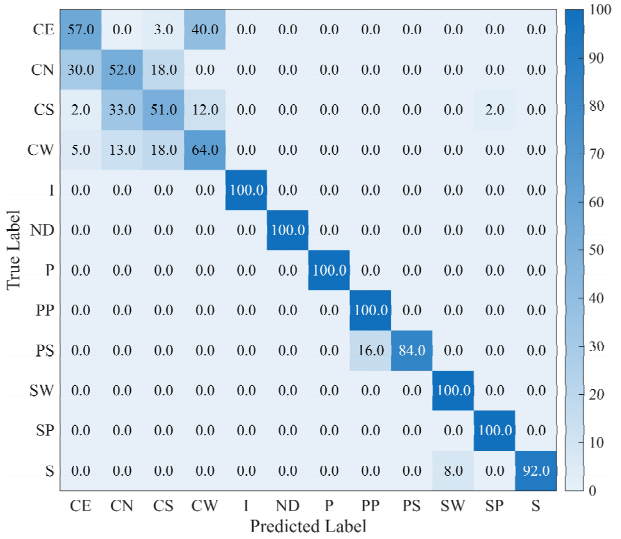


Fig. 3 Normalized confusion matrix (zero-shot)

정답 클래스 y_i에 대한 배타적 대조 손실은 긍정 텍스트와 부정 텍스트 집합 $\boldsymbol{N}(y_i)$를 사용하여 식 (11)과 같이 정의된다. w_{y_i,c^-}는 부정 참조 클래스(c^-)의 반영 비율이며, $\sum_{c^-\in\boldsymbol{N}(y_i)} w_{y_i,c^-}=1$로 정규화한다. 과적합을 최소화하기 위해 학습에 사용된 하이퍼파라미터는 보수적으로 적용하였으며, Table 5와 같다.

$$\boldsymbol{v_i}=\frac{E_I(x_i)}{\| E_I(x_i) \|} \tag{9}$$

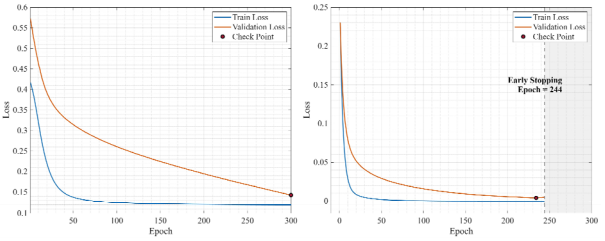
$$z_{i,c}=\exp(\gamma)\left(\boldsymbol{v_i}^{\top}\bar{t}_c\right) \tag{10}$$

$$L_i=-\log\frac{\exp(z_{i,y_i})}{\exp(z_{i,y_i})+\sum_{c^-\in\boldsymbol{N}(y_i)} w_{y_i,c^-}\exp(z_{i,c^-})} \tag{11}$$

학습 데이터셋 구성 시, 대조 학습 목적으로 각 클래스별 부정 텍스트를 Fig. 3의 맑은 날 가상 데이터로 구성된 검증데이터를 기반으로 산출한 zero-shot 검증 혼동 행렬(Confusion matrix)에서 관측된 오분류 양상을 반영하여 설계하였다. 혼동 행렬의 세로축(True label)은 실제 정답 클래스를, 가로축(Predicted label)은 모델이 예측한 클래스를 나타낸다. 행렬 내 각 셀의 값은 실제 클래스 데이터가 해당 예측 클래스로 분류된 비율을 의미하며, 대각선 요소는 정분류율을, 그 외의 요소는 타 클래스와의 혼동

Table 6 Negative reference class composition ratios derived from the zero-shot confusion matrix

Class	Negative reference classes	Composition ratio
CE	CN / CS / CW	10% / 30% / 60%
CN	CE / CS / CW	50% / 30% / 20%
CS	CE / CN / CW	20% / 50% / 30%
CW	CE / CN / CS	20% / 40% / 40%
PS	PP / S	70% / 30%
S	P / SW	30% / 70%



(a) (b)

Fig. 4 Training/validation loss curves for 10-shot(a) and 30-shot(b)

률을 보여준다. 해당 혼동행렬은 부정 텍스트 구성비 결정에만 사용되며, 본 연구의 정량 성능 보고는 별도의 검증 이미지셋에서 수행하였다. 예를 들어, zero-shot 단계에서 타 방위표지들과의 가장 복합적인 혼동이 빈번하게 발생한 CS(Cardinal south)의 부정 텍스트에는 CE(20%), CN(50%), CW(30%)의 긍정 프롬프트 집합을 차용하여 반영하였다. 반면, zero-shot 정확도가 상대적으로 높았던 클래스들은 해당 클래스와 시각적/기하학적으로 상반된 클래스의 긍정 프롬프트를 부정 텍스트로 일괄 적용하여 식별력을 공고히 하였다. 혼동률이 상대적으로 높은 주요 클래스의 부정 텍스트 구성 비율은 Table 6과 같다.

3.2.3 30-shot 튜닝

30-shot 튜닝에서는 기존 10-shot 학습 데이터에 대해 기울기, 밝기, 선명도 조절 및 블러 기법으로 데이터 증강을 적용하여 클래스당 30장으로 학습 표본을 확장하였다. 이때 LoRA 주입 위치, 프롬프트 전략, 하이퍼파라미터 선정, 대조 손실을 포함한 학습 메커니즘은 10-shot 튜닝과 동일하게 유지하여, 학습 데이터 구성이 도메인 적응에 미치는 영향을 분리해 분석할 수 있도록 구성하였다.

Fig. 4는 k-shot에 따른 ViT 모델의 손실 곡선을 비교한 것이다. 샘플 수가 적은 10-shot의 경우 손실값이 완만하게 하강하며 상대적으로 높은 지점에서 수렴하는 반면, 30-shot에서는 학습 초기부터 급격히 하강하여 매우 낮은 손실값에 안정적으로 도달함을 확인하였다. 특히 30-shot 결과에서는 epoch=244에서 조기 종료(early stopping)가 발생하여 최적의 모델이 선정되었다.

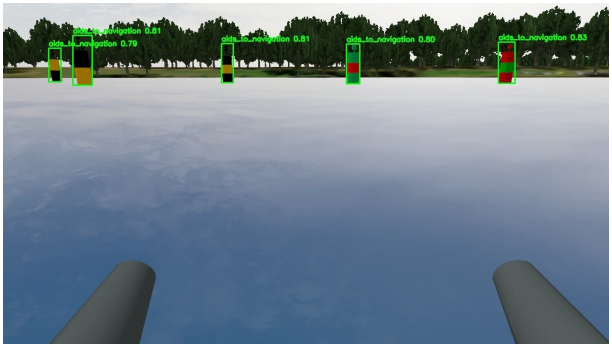


Fig. 5 Example of RoI generation for the online real-time pipeline in the VRX environment

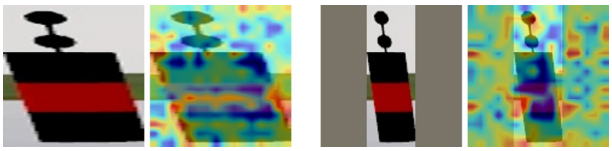
Table 7 Detection post-processing parameters

Parameter	Value
θ_{det}	0.7
θ_{nms}	0.5

3.3 CLIP 실시간 운용 시험 파이프라인 구성

본 절은 4.3절의 실시간 운용 가능성 평가를 수행하기 위해, VRX 환경에서 RoI 생성–전처리–CLIP 분류–모니터링으로 구성된 온라인 시험 파이프라인과 계측 항목을 정의한다. RoI 생성을 위해 사용되는 YOLOv12 기반 모듈은 3.1절에서 설명하는 분류기와는 다른 AtoN 하나의 클래스 검출을 위한 독립적인 네트워크를 통해 분류 입력 RoI를 제공하기 위한 모듈로만 활용된다. 이에 따라 검출기의 정량적 검출 성능 및 실해역 도메인 일반화 평가는 본 논문의 범위에서 제외하였다.

VRX 시뮬레이션의 CLIP 분류 시험에서 카메라 프레임 내 AtoN RoI를 생성하기 위해 활용한 YOLOv12 검출 모듈은 프레임 내 다수의 후보 바운딩 박스를 출력하며, 각 후보는 식 (12)와 같이 정의된다. 여기서, (t_x, t_y)는 박스의 중심 좌표, (t_w, t_h)는 너비와 높이, P_o는 객체의 신뢰도를 의미한다. 검출 결과는 후보 집합 $\boldsymbol{B_0} = \{(b_i, s_i)\}_{i=1}^{M}$ 형태로 얻어지며, b_i는 후보 박스, s_i는 해당 후보의 신뢰도를 의미한다. 오탐을 억제하기 위해 신뢰도 임계값(θ_{det}) 미만의 후보를 제거한 후, 남은 후보들에 대해 NMS(Non-Maximum Suppression)을 적용하여 중복 박스를 제거한다. NMS는 신뢰도(s_i)가 높은 박스를 우선 선택한 뒤, 선택된 박스와의 겹침 정도를 나타내는 IoU (Intersection over Union)가 NMS 임계값(θ_{nms})보다 큰 후보들을 중복 검출로 간주하여 제거하는 방식이다. 최종적으로 바운딩 박스 집합 $\boldsymbol{B} = \{b_1, b_2, ..., b_N\}$를 구성하게 되며 이 과정은 식 (13), (14), (15)과 같다. Table 7은 검출 과정에서 적용되는 임계값을 나타내며, Fig. 5는 VRX 환경에서 AtoN이 탐지된 예시를 보여준다.



(a) (b)

Fig. 6 Comparison of Attention Rollout activation maps, forced resizing(a), letterboxing(b)

$$\boldsymbol{b} = [t_x, t_y, t_w, t_h, P_o] \tag{12}$$

$$\boldsymbol{B}_{conf} = \{(b_i, s_i) \in \boldsymbol{B}_0 \mid s_i \geq \theta_{det}\} \quad (i = 1, ..., N) \tag{13}$$

$$B = NMS(\boldsymbol{B}_{conf}, \theta_{nms}) \tag{14}$$

$$IoU(b_i, b_j) = \frac{|b_i \cap b_j|}{|b_i \cup b_j|} \tag{15}$$

검출된 각 바운딩 박스 b_i에 해당하는 영역은 원본 영상으로부터 RoI 패치 x_i로 크로핑(cropping)되어 분류 단계의 입력으로 전달된다. 이때 AtoN은 두표와 색상띠 패턴 등 상단부의 미세 단서가 분류 결정에 중요한 객체이므로, 두표나 AtoN의 밑단이 박스 경계에서 절단되는 상황을 완화하기 위해 박스 확장 규칙을 적용하였다. 또한 한 프레임에서 다수의 AtoN이 검출되는 경우, 각 RoI 패치에 대해 독립적으로 분류를 수행하게 된다.

CLIP 기반 분류기는 고정된 입력 해상도를 요구하므로 입력 크기 변환이 필수적이며, 종횡비 보존을 위해 padding 기반의 레터박싱 기법이 활용될 수 있다. 그러나 Attention rollout 기반 사례 분석(Fig. 6)에서 레터박싱은 이미지 영역과 패딩 영역 사이의 급격한 픽셀 변화로 인위적 경계가 형성되고, 이로 인해 활성화가 객체가 아닌 경계 영역으로 유도되는 경향을 확인하였다. 반면 강제 리사이징(Forced resizing)은 기하학적 왜곡을 수반하지만, 입력 텐서에서 객체가 차지하는 유효 픽셀 비중을 증가시켜 배경 영향을 줄이고 AtoN의 색상·형상 단서에 집중하도록 유도할 수 있다. 이에 따라 본 연구에서는 CLIP 입력에 대해 RoI 패치를 목표 해상도(224X224)로 강제 리사이징하는 전처리를 채택하였다. 한편 비교군인 YOLOv12 분류기는 입력 크기를 모델 설정에 맞추기 위한 리사이즈/크롭 등의 전처리가 추론 파이프라인 내부에서 자동으로 수행된다. 따라서 YOLOv12 분류기에서는 RoI 패치를 그대로 입력으로 전달하여 최종 입력 정규화는 해당 모델의 기본 전처리 절차에 따라 처리되도록 설정하였다.

실시간 운용 관점에서, 최종 출력은 각 객체에 대해 $(b_i, \hat{c}_i, s_i)$ 형태로 정의할 수 있으며, 여기서 $\hat{c}_i$는 예측 클래스, s_i는 검출 신뢰도와 분류 신뢰도를 결합한 최종 점수로 식 (16)과 같이 설정할 수 있다. 또한 ROS2 기반 시스템에서 CLIP 분류기의 처리 시간을 측정하여 해당 분류 방식의 실시간 적용 가능성을 4.3절에서 검토한다.

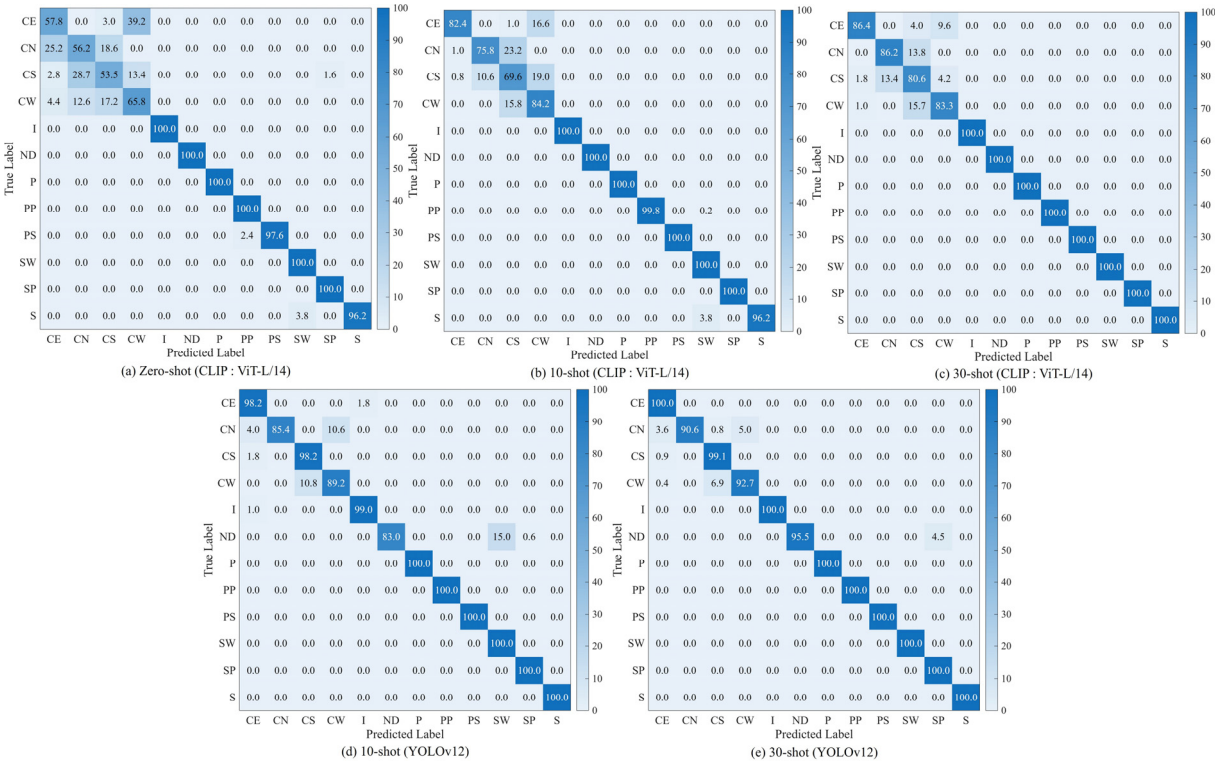


Fig. 7 Normalized confusion matrices under clear conditions in the VRX

$$s_i = P_o \cdot \max_c P(c|x_i) \quad (16)$$

4. 검증 결과 및 분석

4.1 시뮬레이션 환경 검증

4.1.1 평가 지표

본 연구에서는 분류 결과를 혼동 행렬(M)으로 정리하여 성능을 평가하였다. $M_{n,m}$는 정답이 n인 샘플이 m으로 예측된 개수를 의미한다. 정확도(Accuracy)는 전체 샘플 중 정분류 비율로 정의된다. 또한 각 클래스 c에 대해 one-vs-rest 관점에서 정밀도(Precision), 재현율(Recall), F1-score를 계산하였다. 본 연구에서는 클래스별 F1-score를 단순 평균한 Macro-F1을 핵심 지표로 사용하였다. Macro-F1은 클래스 불균형의 영향을 완화하여 미세 분류 성능을 보수적으로 평가하는 지표이다.

4.1.2 데이터 효율성 평가

본 연구는 데이터 확보가 어려운 해상 환경이라는 실무적 제약 상황을 상정하였으며, 이러한 저자원 환경에서 VLM이 기존 지도학습 모델의 한계를 보완할 수 있는 실질적 대안이 될 수 있는지 검증하는 데 목적이 있다. 맑은 날 조건의 검증용 가상 데이터셋에 대해 제안하는 CLIP 모델과 베이스라인인 YOLOv12의 분류

Table 8 Classification results under clear conditions in the VRX

Model	n-Shot	Accuracy	Precision	Recall	Macro-F1
YOLOv12	10	0.961	0.965	0.961	0.962
	30	0.981	0.982	0.981	0.981
CLIP (ViT-L/14)	zero	0.856	0.856	0.856	0.855
	10	0.923	0.930	0.923	0.925
	30	0.947	0.952	0.947	0.949

성능을 비교 분석하였다. Fig. 7은 학습 데이터 양 변화에 따른 정규화된 혼동 행렬을 나타내며, Table 8은 이를 바탕으로 산출한 정확도, 정밀도 및 Macro-F1 score를 요약한다.

실험 결과에서 가장 주목할 점은 도메인에 특화된 학습 데이터 없이도 분류 작업을 수행할 수 있는 CLIP의 zero-shot 성능이다. YOLOv12 분류기는 지도학습 기반이므로 AtoN 라벨 데이터가 주어지지 않으면 분류기가 성립하지 않는 반면, CLIP은 도메인 학습 데이터가 전무한 zero-shot 조건에서도 0.856의 정확도와 0.855의 Macro-F1 score를 기록하며 초기 운용 단계에서 유의미한 성능을 보였다. 특히 방위표지(CE, CN, CS, CW)를 제외한 다수 클래스에서는 높은 재현율을 달성하였다. 이는 규정 기반 속성을 텍스트 프롬프트로 명시함으로써, 사전학습된 시각-언어 정렬 지식을 라벨 부족 환경에서 직접 활용할 수 있기 때문이다.

반면 시각적 형태 및 색상 조합이 유사한 4가지 방위표지 간의 미세한 차이를 구분하는 데 있어 약 50~60%대의 재현율을 보이며 한계를 드러내기도 했다. 그러나 이러한 한계는 소량의 데이

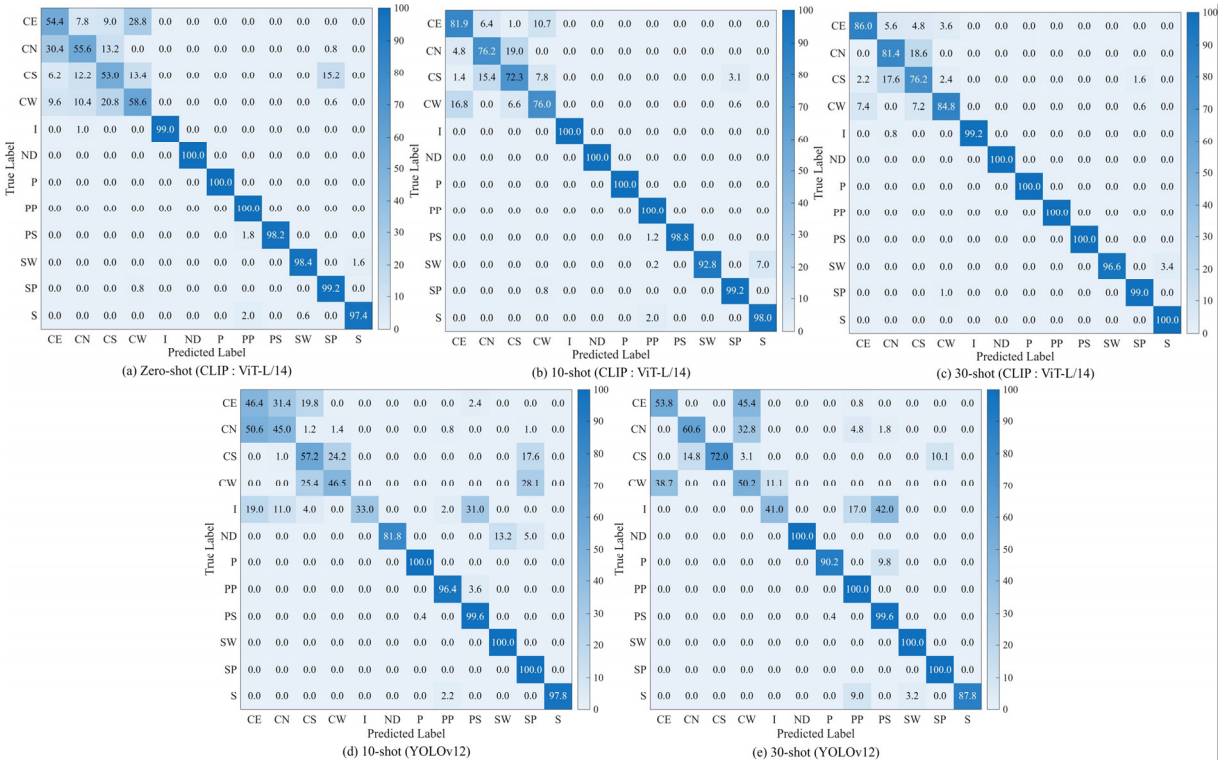


Fig. 8 Normalized confusion matrices under foggy conditions in the VRX

터 주입만으로 빠르게 완화되었다. 클래스당 10장의 학습 샘플만을 사용한 10-shot 조건에서 CLIP 모델의 정확도는 0.923으로 급격히 상승하였으며, 방위표지 간의 혼동 역시 크게 개선되었다. 이는 VLM이 사전 학습된 지식을 바탕으로 소량의 도메인 힌트를 통해 핵심 혼동 패턴을 효율적으로 보정할 수 있음을 입증한다.

30-shot 조건에서 YOLOv12는 가상 데이터 분포 내에서 매우 높은 성능을 보였으며, 이는 합성 데이터의 시각적 규칙성이 지도학습 분류기에 유리하게 작용할 수 있음을 드러낸다. 반면에, CLIP 역시 0.947로 높은 성능을 보이며 베이스라인에 근접하였다. 합성 데이터에서의 높은 성능이 해무와 같은 시각적 방해 요소가 존재하는 환경에서도 유지되는지 확인하기 위해, 다음 절에서는 시각적 방해 요인이 존재하는 조건에서 분류 강건성을 추가로 검증한다.

4.1.3 환경 강건성 평가

해무 조건은 색상 대비를 감소시키고 경계를 흐리게 만들어 AtoN의 규정 기반 미세 단서가 쉽게 소실되는 환경이다. 이에 본 절에서는 맑은 날의 학습 데이터로 학습된 분류 모델을 해무가 적용된 VRX 가상 검증 데이터에 그대로 적용하여, 환경 변화에 대한 강건성을 평가하였다. Fig. 8은 각 모델의 정규화 혼동 행렬을, Table 9는 정량 지표를 요약한다.

해무 조건에서 CLIP 기반 분류기는 zero-shot에서도 정확도 0.843, Macro-F1 0.842을 기록하여 일정 수준의 성능을 유지하는 모습을 보였다. 맑은 날 대비 성능이 소폭 하락하였으며, 성능

Table 9 Classification results under foggy conditions in the VRX

Model	k-Shot	Accuracy	Precision	Recall	Macro-F1
YOLOv12	10	0.740	0.756	0.740	0.746
	30	0.792	0.804	0.792	0.798
CLIP (ViT-L/14)	zero	0.843	0.845	0.843	0.842
	10	0.911	0.918	0.911	0.914
	30	0.934	0.941	0.934	0.937

저하는 방위표지 간 상호 혼동으로 집중되었으나, 맑은 날 데이터만으로 10-shot 및 30-shot LoRA 튜닝을 적용하면 방위표지군의 혼동이 완화되며 Macro-F1이 각각 0.914, 0.937로 향상되었다. 이는 해무로 인해 미세 단서가 약화되는 조건에서도 성능 저하를 제한하거나, 소량의 샘플을 통한 튜닝만으로 핵심 혼동 패턴을 효율적으로 보정할 수 있음을 보여준다.

반면 YOLOv12 분류기는 해무 조건에서 성능 저하가 뚜렷하였다. 10-shot 조건에서 정확도 0.740, Macro-F1 0.746로 감소하였고, 30-shot으로 학습 데이터를 확장하더라도 정확도 0.792, Macro-F1 0.798 수준에 머물렀다. 특히 방위표지 간 상호 혼동이 크게 증가하고, 일부 클래스는 특정 표지군으로 오분류가 집중되는 경향이 관측되었다.

Table 10에서 시현된 CLIP의 높은 성능 유지율(98.5~98.8%)은 VLM 아키텍처가 갖는 고유의 강건성에 기인한 것으로 분석된다. 반면, 지도 학습 기반의 YOLOv12는 해무 환경에서 약 20%의 성능 손실을 기록하였는데, 이는 소량의 학습 데이터로 인해

Table 10 Performance degradation summary from clear to foggy conditions in the VRX

Model	k-Shot	Macro F1 (Clear)	Macro F1 (Fog)	△Macro-F1	Retention
YOLOv12	10	0.962	0.746	-0.216	77.5%
	30	0.981	0.798	-0.183	81.3%
CLIP (ViT-L/14)	zero	0.855	0.842	-0.013	98.5%
	10	0.925	0.914	-0.011	98.8%
	30	0.949	0.937	-0.012	98.7%

특정 관측 조건에 과적합되기 쉬운 지도 학습 모델의 도메인 적응 취약성을 시사한다. 특히 CLIP은 LoRA 기반의 파라미터 효율적 미세 조정을 통해 사전 학습된 범용적 시각 표현력을 보존하는 동시에, 텍스트 프롬프트에 명시된 AtoN 규정 속성과 이미지 간의 의미론적 정렬을 수행한다. 이러한 특성은 픽셀 단위의 시각 정보가 손상되는 조건에서도 객체의 본질적 단서를 유지하게 함으로써, 제안 기법이 기존 지도 학습 분류기 대비 우월한 환경 강건성을 확보하였음을 실증한다.

4.2 시뮬레이션-실제 환경 일반화 검토

VRX 가상 데이터 기반으로 학습된 CLIP 분류기가 실해역에서 촬영된 공개된 해상표지 이미지에서 어떤 시각 단서에 근거하여 분류 결정을 내리는지 확인하기 위해 Attention rollout을 적용하였다. 실해역 공개 이미지로는 PP/PS를 충분히 확보하지 못해, 실해역 평가는 10개의 클래스 20장씩 총 200장에 한정하였다. 본 연구에서는 실제 해상 환경에서의 성능을 평가하되, 검출기 성능 변동이 분류 성능 평가를 혼합하지 않도록 분류기 자체의 도메인 일반화 성능을 분리하여 분석한다. 이를 위해 실해역 이미지에서 AtoN이 포함된 수동으로 크롭한 RoI-given 패치에 대해 YOLOv12 분류기와 CLIP 기반 분류기의 성능을 비교하였다.

Attention rollout은 CLIP의 이미지 인코더인 ViT-L/14의 내부 Self-attention을 계층적으로 누적하여 토큰 간 영향 전이를 계산한다. 본 연구에서는 기존 rollout만 사용하지 않고, top-1과 top-2 예측 클래스의 마진(식 (17))에 대한 gradient를 이용해 각 층의 head-평균 Attention($\Phi^{(l)}$)을 가중하였다. $\Phi^{(l)}$에 대해 $G^{(l)} = \partial J / \partial \Phi^{(l)}$를 계산하고, 양(+)의 기여만 반영하기 위해 식 (18)을 통해 가중 Attention을 구성하였다. $\tilde{\Phi}^{(l)}$은 층 방향으로 누적되어 relevance 행렬을 얻을 수 있게 되고, 클래스 토큰이 패치 토큰에 할당한 성분을 2D로 재배열해 heatmap을 생성하였다.

$$J(x_i) = z_{i,\hat{c}} - z_{i,c^{(2)}} \tag{17}$$

$$\tilde{\Phi}^{(l)} = ReLU\left(G^{(l)}\right) \odot \Phi^{(l)} \tag{18}$$

Fig. 9는 실해역 테스트 이미지의 RoI 패치에 대해, 제안한 변형 Attention rollout을 적용한 시각화 결과를 제시하며, 사용된 이미지의 출처 및 라이선스는 부록의 Table 13에 정리하였다. 녹색 테두리는 정분류 사례, 적색 테두리는 오분류 사례이며, 오분류 사례의 heatmap은 모델이 잘못된 예측을 선택하는 과정에서 상대적으로 크게 기여한 영역을 나타낸다.

Table 11 Classification results for real images

Model	k-Shot	Accuracy	Precision	Recall	Macro-F1
YOLOv12	30	0.395	0.401	0.395	0.386
CLIP	30	0.860	0.871	0.860	0.862

정분류 사례에서는 모델의 활성화가 대체로 AtoN의 규정 기반 핵심 단서에 대응되는 영역에 형성되는 경향이 관찰되었다. 예를 들어, 고립 장애물 표지(I)는 구형 두표 및 몸체 패턴 영역, 신위험물 표지(ND)는 '+'형 두표와 색상 패턴, 특수표지(SP)는 'X'형 두표와 몸체, 안전 수역 표지(SW)는 상부구조 및 표지 본체의 주요 구성요소에 활성화가 집중되었다. 이는 시뮬레이션 기반 학습만으로도 실제 환경에서 AtoN을 구분하는 구조적 단서를 일정 수준 활용하고 있음을 정성적으로 뒷받침한다.

반면 오분류 사례는 주로 방위표지군에서 나타났으며, 정분류 사례와 비교할 때 띠 패턴의 배치와 같은 구분 단서에 충분히 집중하지 못하는 것을 확인하였다. AtoN의 가장 하단을 배경과 헷갈리거나, 구조물에 있는 부착물(이끼 등)으로 인해 미세 단서가 불명확해질 수 있으며, 이러한 조건에서 방위표지군 내 상호 혼동이 증가하는 경향이 heatmap에서도 관찰되었다.

Fig. 10과 Fig. 11은 각각 실제 해상 AtoN 이미지에 대해 30-shot의 YOLOv12와 CLIP 분류기의 정규화 혼동 행렬을 나타낸다. CLIP의 경우는 대부분의 클래스에서 대각 성분이 우세하여, 제한된 가상 데이터 튜닝만으로도 클래스 간 구분이 안정적으로 유지됨을 확인하였다. 하지만, YOLOv12는 대각 행렬의 우세함이 CLIP에 비해 확연히 떨어지는 것을 확인하였다.

Table 11은 동일한 실제 이미지 기반 RoI-given 입력을 YOLOv12 기반 분류기와 CLIP 기반 분류기에 공통으로 적용하여 정량적인 성능을 비교한 결과이다. 이때 두 분류기는 모두 VRX 맑은 날 합성 데이터에서 30-shot으로 학습된 설정에서 도출된 최적 성능 체크 포인트를 사용하여 분류 작업을 수행하였다. 또한 추론 단계의 입력 전처리 및 클래스 라벨 매핑은 학습 시 설정과 동일하게 유지하였다. YOLOv12 분류기가 실해역 테스트에서 0.386의 낮은 Macro-F1을 기록한 것은 합성 데이터 기반 학습 분포와 실환경 이미지 간의 극심한 도메인 간극을 시사한다. 가상 환경에서 생성된 AtoN은 표면이 균일하고 시각적 단서가 명확한 반면, 실제 해상 객체는 노후에 따른 부식, 조류 부착, 그리고 태양광에 의한 강한 반사 및 역광 등 지도 학습 모델이 학습 단계에서 경험하지 못한 비정형적 노이즈를 다수 포함한다. 이러한 데이터 분포의 변화에 취약하여 도메인 일반화에 실패한 것으로 분석된다. 반면, CLIP은 텍스트 프롬프트를 통한 의미론적 속성 주입을 통해 시각적 오염을 관통하는 객체의 본질적 형상과 색상 조합을 식별해냈다. 이는 실해역 데이터 확보가 어려운 초기 운

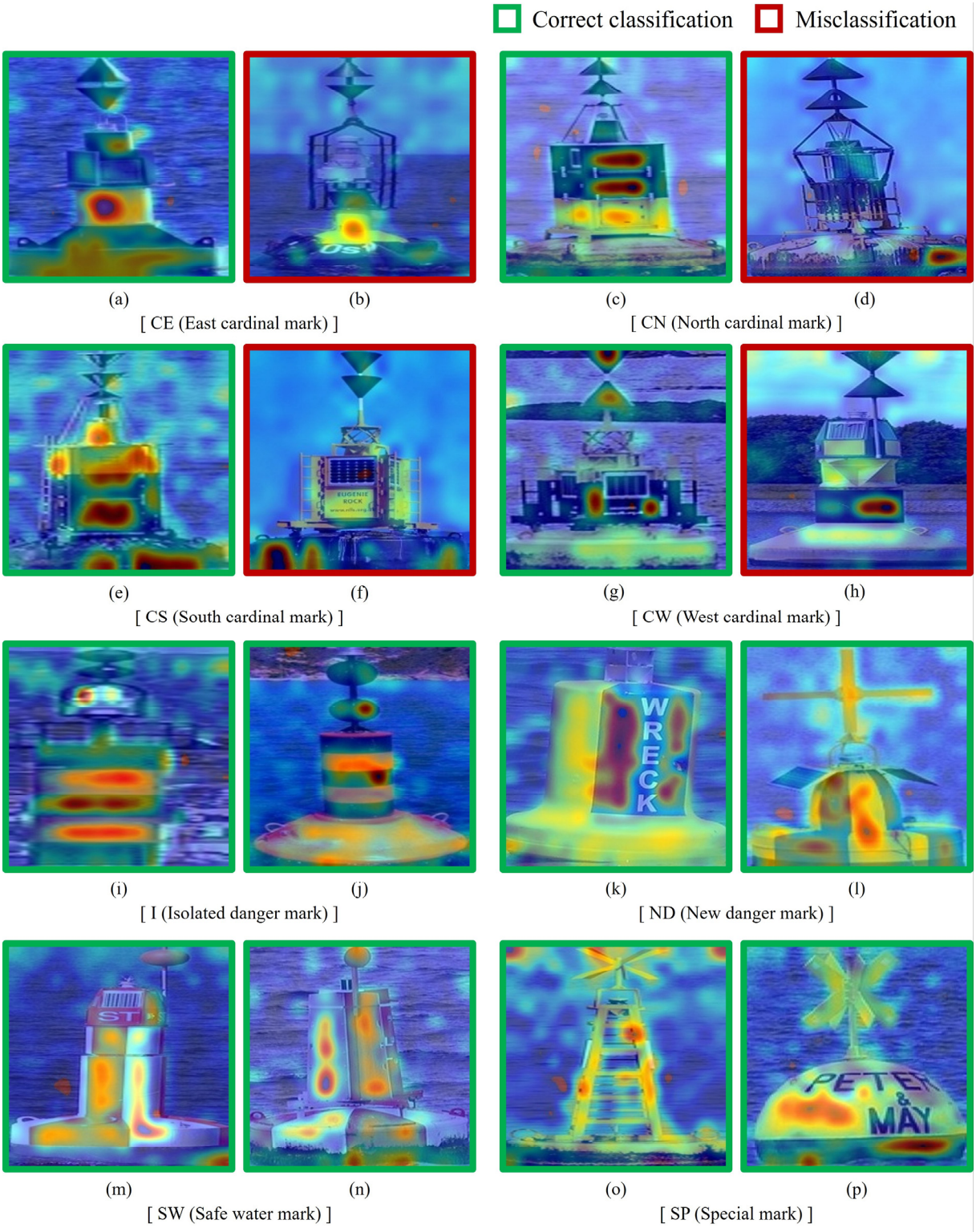


Fig. 9 Gradient-weighted Attention rollout visualizations of the CLIP-based AtoN classifier on real-sea RoI patches

용 환경에서, 가상 데이터만으로 튜닝된 VLM이 기존 지도 학습 모델보다 월등한 실질적 대안이 될 수 있음을 입증하는 결과이다. 단, 본 결과는 RoI-given 조건에서의 분류 성능 비교이며, 검출 오차가 결합되는 end-to-end 운용 성능의 열화 양상은 별도의 통합 평가를 통해 정량화할 필요가 있다.

4.3 실시간 운용 가능성 평가

본 연구에서는 3.3절에서 설계한 RoI 기반 온라인 처리 구조를 기반하여 CLIP 분류 모듈의 연산 지연을 계측하여 실시간 적용 가능성을 논의한다. 분류 결과와 처리 지연을 즉시 확인할 수

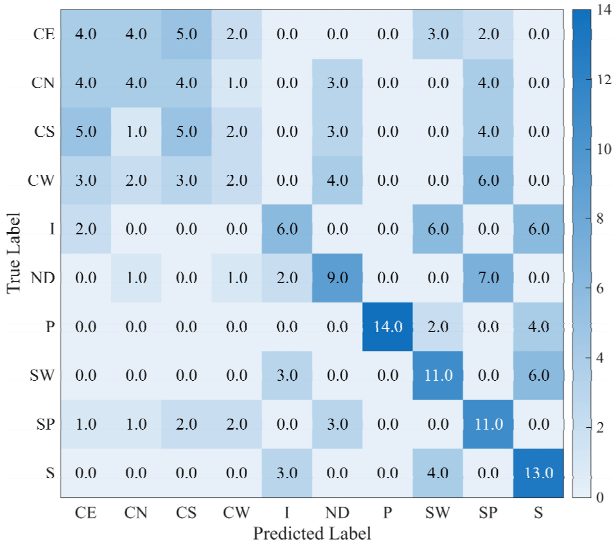


Fig. 10 Normalized confusion matrix of the YOLOv12-based classifier for real-world image classification

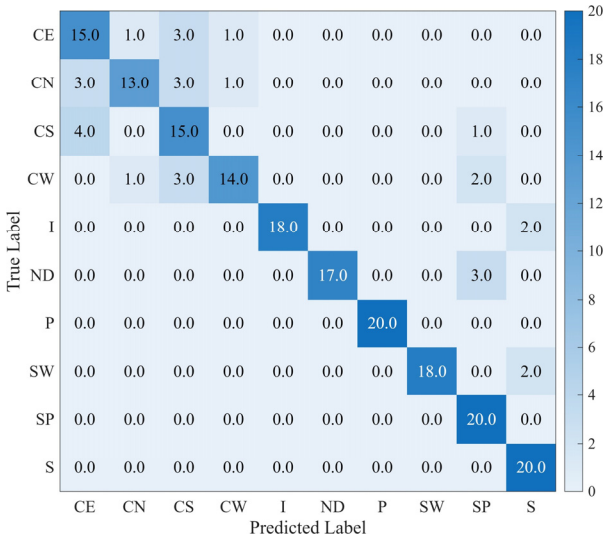


Fig. 11 Normalized confusion matrix of the CLIP-based classifier for real-world image classification

있도록 Fig. 12와 같이 웹 기반 대시보드를 별도로 구현하였다. 해당 대시보드는 YOLO 검출 영상, CLIP 입력 전처리 결과, Attention rollout 시각화, 처리 지연시간 등을 단일 화면에서 갱신하도록 구성하였으며, VRX 시뮬레이션 환경에서 실제로 파이프라인을 구동하여 지연시간 및 분류 결과 로그를 수집하였다. 이와 같은 운용자 관점의 시각적 모니터링 및 계측 절차는 CLIP 분류 모듈의 지연 특성과 온라인 적용 가능성을 논의하기 위한 실험적 근거를 제공한다.

Table 12는 VRX 시뮬레이션 환경에서 RoI 기반 처리 파이프라인을 구성한 뒤, CLIP 분류 모듈의 추론 지연을 단계별로 분해하여 계측한 결과를 요약한 것이다. 본 계측은 분류 모듈의 연산 지연 규모와 지연 기여도를 정량적으로 제시하기 위한 프로파일링 시험으로 수행되었다. 견시 지원에서는 AtoN이 항상 고속 갱신이 필요한 타겟이 아니며, 본 연구는 CLIP 분류 1회 비용이 실

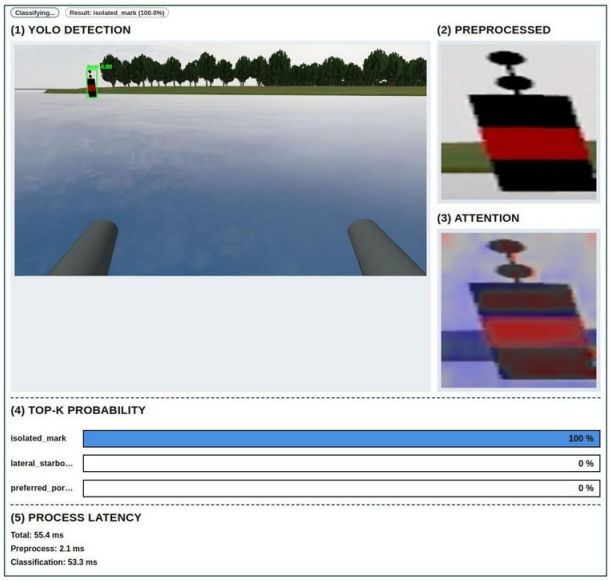


Fig. 12 Real-time monitoring dashboard in the VRX

Table 12 Topic-based processing latency measured in the VRX

Metric	Mean	95% within	99% within
Total	59.59 ms	63.85 ms	100.07 ms
Detection	1.15 ms	1.33 ms	1.47 ms
Preprocess	3.14 ms	6.72 ms	44.26 ms
Classification	55.30 ms	58.86 ms	59.69 ms

시간 운용 범위에 들어오는지의 프로파일링에 초점을 둔다.

단일 AtoN RoI에 대해 전체 연산 지연은 평균 약 60 ms로 측정되었으며, RoI 검출과 전처리 단계는 각각 평균 1.15 ms, 3.14 ms로 지연 기여가 상대적으로 작았다. 반면 CLIP 기반 분류 단계는 평균 55.30 ms(99% within 59.69 ms)로 전체 지연의 대부분을 차지하였는데, 이는 전체 처리 지연이 주로 CLIP 인코딩 및 텍스트-이미지 유사도 계산 시간에 의해 결정적으로 좌우됨을 의미한다. 또한 이는 CLIP 분류 모듈이 단일 RoI 기준 약 55 ms($\approx$ 18 Hz 수준)의 처리율로 동작함을 보여주며, 본 실험 조건에서 온라인 적용을 위한 처리율을 확보했음을 시사한다.

다만 지연 분석은 ROS2 토픽 기반으로 계측된 연산 지연에 한정되며, 카메라 취득 및 데이터 저장/전송, 노드 간 통신 지연을 포함한 전체 end-to-end 지연 평가는 추후 통합 계측을 통해 보완할 필요가 있다.

5. 결 론

본 연구는 MASS의 연안 및 항만 운용 시 필수적인 AtoN의 미세 분류를 위해, RoI 기반의 CLIP 분류 기법을 제안하였고, 실험을 통해 데이터 효율성, 환경 강건성 및 온라인 적용 가능성을 함께 검토하였다. VRX 가상 데이터 실험에서 CLIP 분류기는 zero-shot 조건에서도 유의미한 분류 성능(정확도 0.856, Macro-F1 0.855)을

보였으며, 10-shot 및 30-shot의 소량 데이터 주입을 통해 성능이 빠르게 향상됨을 확인하였다. 또한 해무 조건에서도 Macro-F1 유지율이 98.5~98.7%로 나타나, 환경 변화에 따른 성능 저하가 상대적으로 작게 관찰되었다. 특히 분류기의 도메인 일반화 성능을 명확히 보기 위해 수행한 실해역 RoI-given 테스트에서도 CLIP 30-shot은 정확도 0.860, Macro-F1 0.862를 기록하여, 제한된 가상 데이터 튜닝만으로도 실환경에서의 분류 성능이 유지될 수 있음을 보이는 예비적 근거를 제시하였다. 다만, 본 실험에서 활용된 실해역 이미지의 모수가 제한적이므로, 향후 다양한 해역 및 기상 조건을 포함하는 대규모 실해역 데이터셋을 확충하여 검증의 통계적 신뢰성을 더욱 강화할 필요가 있다.

VRX 시뮬레이션을 통한 ROS2 토픽 기반 지연 계측에서 단일 RoI 처리 시간은 평균 약 60 ms였고, 이 중 CLIP 분류 단계는 평균 약 55 ms로 측정되어 RoI 입력 기준의 분류 모듈 온라인 적용 가능성을 뒷받침한다. 본 연구의 실해역 평가는 분류기의 일반화 성능을 분리 검증하기 위한 RoI-given 설정이므로, 검출-분류 통합 운용 성능을 직접적으로 대변하지는 않는다. 또한 실시간 운용 가능성 검토를 위해 수행한 파이프라인에서 AtoN 단일 클래스 검출 실패는 검출-분류 통합 운용 성능을 좌우한다는 한계를 가진다. 향후 연구에서는 확충된 실해역 데이터를 바탕으로 실제 환경의 원본 카메라 프레임에서 다중 AtoN을 검출 및 분류하고, IALA 규정 기반의 표지 조합 관계로부터 항행 의미를 추론하는 장면 수준 의미론적 추론을 통해 견시 지원의 end-to-end 운용으로 확장할 계획이다.

후 기

본 연구는 한국해양과학기술원 부설 선박해양플랜트연구소의 기본연구사업 “항계 내 자율운항선박 운용 안전성·적합성 평가 기술 개발(3/5) [PES5880]”에 의해 수행되었습니다. 이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2022R1I1A3064596).

References

Beaumont, R., 2022. Large scale OPENCLIP: L/14, H/14 and G/14 trained on LAION-2B [Online] (Uptated 15 September 2022) Available at: https://laion.ai/blog/large-openclip/ [Accessed 23 March 2026].

Han, X., Pan, M., Ge, H., Li, S., Hu, J., Zhao, L. and Li, Y., 2021. Multilabel video classification model of navigation mark's lights based on deep learning. *Computational Intelligence and Neuroscience,* 6794202.

International Maritime Organization (IMO), 1972. Convention on the international regulations for preventing collisions at sea, 1972 (COLREGs).

International Maritime Organization (IMO), 1995. Performance standards for electronic chart display and information systems (ECDIS). Resolution A.817(19).

International Maritime Organization (IMO), 2022. ECDIS-Guidance for good practice. MSC.1/Circ.1503/Rev.2.

Choi, J., Park, J., Kang, M., Kim, H., Youn, W., 2023. Multiple PDAF algorithm for estimation states multiple of the ships. *Journal of the Society of Naval Architects of Korea,* 60(4), pp.248-255.

Kaur, P., Aziz, A., Jain, D., Patel, H., Hirokawa, J. and Townsend, L., 2022. Sea situational awareness (SeaSAw) dataset. 2022 Institute of electrical and electronics engineers/computer vision foundation conference on computer vision and pattern recognition workshops, *New Orleans*, LA, USA, 19-24 June 2022.

Nam, G.W., Roh, M.I., Lee, H.W. and Lee, W.J., 2021. Classification of ship images for autonomous ships using deep learning. *Korean Journal of Computational Design and Engineering*, 26(2), pp.144-153.

Park, J.-H., Roh, M.-I., Lee, H.-W., Jo, Y.-M., Ha, J., Son, N.-S., 2024. Multi-vessel target tracking with camera fusion for unmanned surface vehicles. *International Journal of Naval Architecture and Ocean Engineering,* 16, 100608.

Pan, M., Liu, Y., Cao, J., Li, Y., Li, C. and Chen, S., 2020. Visual recognition based on deep learning for navigation mark classification. *Institute of Electrical and Electronics Engineers Access,* 8, pp.32767-32775.

Perera, L.P., 2019. Deep learning toward autonomous ship navigation and possible COLREGs failures. *Journal of Offshore Mechanics and Arctic Engineering,* 142(3), 031302.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., 2021. Learning transferable visual models from natural language supervision. 38th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, pp.8748-8763.

Schöller, F.E.T., Nalpantidis, L. and Blanke, M., 2022. Buoy light pattern classification for autonomous ship navigation using recurrent neural networks. *Institute of Electrical and Electronics Engineers Transactions on Intelligent Transportation Systems,* 23(7), pp.9455-9465.

Sim, Y. and Chae, C.J., 2025. Enhancing the function of the aids to navigation by practical usage of the deep learning algorithm. *The Journal of Navigation,* 77(3), pp.347-358.

Tian, Y., Ye, Q. and Doermann, D., 2025. YOLOv12: Attention-centric real-time object detectors.

Wang, H., Shi, J., Karimian, H., Liu, F. and Wang, F., 2024. YOLOSAR-Lite: A lightweight framework for real-time ship detections in SAR imagery. *International Journal of Digital Earth,* 17(1), 2315729.

Yeo, I.-C., Roh, M.-I., Kim, Y.-S., Kim, H.-Y., Ahn, D.-H., Son, N.-S., 2025. A localization method of nearby ships based on 3D object detection using a camera. *International Journal of Naval Architecture and Ocean Engineering,* 17, 100705.

Zhen, R., Ye, Y., Chen, X. and Xu, L., 2023. A novel intelligent detection algorithm of aids to navigation based on improved YOLOv4. *Journal of Marine Science and Engineering,* 11(2), 452.

Authorship Contribution Statement

Sun-Hyuck Im: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Jungin-Hyeok Seo**: Data curation, Formal analysis, Validation; **Si-Won Kim**: Methodology, Validation; **Seong-Hyeon**: Data curation, Formal analysis, Methodoloy; **Yeon-Soo Kim**: Data curation, Visualization; **Hyun-Jae Jo**: Investigation, Formal analysis; **Jong-Yong Park**: Funding acqusition, Supervision, Writing – review & editing.

부 록

Table 13 Attribution and license information for the images used in Fig. 10 and applied modifications

Image		Author	Source	License	Modification
Fig. 9	(a)	–	Safe Transport Victoria	CC BY 4.0	Crop, Resize, Heatmap
	(b)	Reinhard Kraasch	Wikimedia Commons		
	(c)	David Dixon			
	(d)	Canthusus			
	(e)	David Dixon			
	(f)	Toby Speight			
	(g)	Peter Moore			
	(h)	Ein Dahmer			
	(i)	Waterloooo			
	(j)	Mobilis			
	(k)	ITookSomePhotos			
	(l)	–	Safe Transport Victoria		
	(m)	Wasser	Wikimedia Commons		
	(n)	Rab Farrow			
	(o)	David Dixon			
	(p)	Lan Paterson			