Journal of the KNST

2026; Vol.9, No.1; pp.019-024

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2026.3.9.1.19

Received: 2026/01/19
Revised: 2026/02/02
Accepted: 2026/03/09
Published: 2026/03/31

*Corresponding Author:
Kayeon Kim
Unmanned Platform R&D Center, LIG Nex1, 13494 231, Pangyoyeok-ro, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea
Tel: +82-31-8038-0768
E-mail: kayeon.kim@lignex1.com

수중 기뢰 탐지 및 분류를 위한 Vision-Language Model 분석

Analysis of Vision-Language Models for Underwater Mine Detection and Classification

김가연[1*], 장현배[2], 박석준[3], 김윤호[2]

[1]LIG넥스원 무인플랫폼연구소 연구원
[2]LIG넥스원 무인플랫폼연구소 선임연구원
[3]LIG넥스원 무인플랫폼연구소 수석연구원

Kayeon Kim[1*], Hyunbae Chang[2], Seokjoon Park[3], Yoonho Kim[2]

[1]Researcher, Unmanned Platform R&D Center, LIG Nex1
[2]Senior researcher, Unmanned Platform R&D Center, LIG Nex1
[3]Chief researcher, Unmanned Platform R&D Center, LIG Nex1

Abstract

수중 기뢰는 함선·선박 파괴와 해상 봉쇄 등 전략적 역할을 수행하므로 기뢰 탐지 연구가 필수적이다. 이에 따라, 급변하는 해양 환경, 다양한 기뢰 형태, 전시 상황에서의 작전 변화에 적응할 수 있는 모델이 필요하다. Vision-Language Model(VLM)은 이미지와 텍스트를 동시에 처리할 수 있는 모델로, 변화가 잦은 수중 환경에 적합하다. 본 논문은 최신 VLM을 분석하고, 이를 수중 기뢰 탐지·분류에 적용하는 방안을 모색한다.

Underwater mines play a strategic role in destroying ships and vessels and enforcing maritime blockades. Therefore, mine detection research is essential. Accordingly, a model that can adapt to rapidly changing marine environments, various mine types, and operational shifts in wartime is required. Vision-language models (VLMs) can process both images and texts simultaneously, making them well-suited for the frequently changing underwater environment. This paper analyzes the latest VLMs and explores methods to apply them to underwater mine detection and classification.

Keywords

수중 기뢰 탐지(Underwater Mine Detection), 비전-언어 모델(Vision-Language Model), 멀티모달 데이터셋(Multimodal Dataset), 이미지 분류(Image Classification), 오픈-월드 객체 탐지(Open-World Object Detection)

Acknowledgement

이 논문은 2023년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(No. KRIT-CT-23-035-03, AI기반 수중 기뢰 탐지 기술(기뢰탐지용 무인잠수정 군집 운용 기술)).

1. 서론

기뢰는 수중에 부설되어 함선·선박을 파괴하는 무기로, 항구나 해상 봉쇄, 경로 차단 등 전략적 목적으로 활용된다. 전시 상황에서 기뢰의 존재 여부를 신속하고 정확하게 파악하는 것은 작전 성공 여부를 좌우할 정도로 중요한 과제이다. 최근 딥러닝이 수중 물체 탐지 분야에 폭넓게 적용되면서, 소나 영상으로부터 기뢰를 식별하는 모델에 대한 연구가 늘어나고 있다. 기존의 이미지 기반 접근 방식은 대체로 사전에 정의된 제한된 클래스만을 인식할 수 있다는 제약을 가지고 있다[1]. 기뢰는 다양한 유형으로 존재하며, 수중 환경 조건과 해저 지형에 따라 매몰 상태와 노출 양상이 서로 다르게 나타난다. 더욱이 전시 상황에서 작전이 매번 동일하게 전개되지 않고 지속적으로 변화하기에, 센서 운용 방식, 해역 특성, 수중 지형, 기뢰 형태의 노출 양상 등 기뢰 탐지 조건 전반에 영향을 미친다. 이로 인해 작전 변동성에 적응할 수 있는 모델이 필요하다. 이러한 한계를 극복할 수 있는 방안으로 Vision-Language Model(VLM)을 주목한다. VLM은 이미지와 텍스트를 동시에 처리할 수 있는 모델이다. 이는 사전에 정의된 고정 클래스에 얽매이지 않은 open-world 탐지를 가능하게 하여, 동일 카테고리를 가진 다양한 형태의 기뢰를 효과적으로 인식하도록 한다. 도메인 특화

지식을 프롬프트 형태로 구성하여 모델에 입력함으로써, 새로운 기뢰 유형이나 급변하는 해양 환경, 다양한 작전 환경에 대한 적응성을 향상시킨다. 본 논문에서는 VLM의 최근 연구 동향을 살펴보고, 이를 수중 기뢰 탐지 및 분류에 새로운 가능성을 제공하는지 분석한다. 더 나아가 수중 기뢰 탐지 및 분류에 VLM을 적용할 수 있는 방안에 대해 탐색한다.

2. VLM 연구 분석

Vision-Language Model(VLM)[1-4]은 이미지와 텍스트를 동시에 이해하고 처리할 수 있는 모델이다. 이미지와 텍스트를 각각 해석하기 위해 image encoder, text encoder를 사용하고, 추출된 특징들을 결합함으로써 오픈 시나리오에서의 탐지를 수행할 수 있다. 이를 통해, VLM은 image classification, object detection, image captioning, image retrieval로 활용될 수 있다. 특히 이미지와 텍스트 정보를 모두 활용하는 객체 탐지 및 분류 능력은 수중 기뢰 탐지에 직접 적용할 수 있다. VLM을 기뢰 이미지 분류 및 탐지에 적용하면, 정보의 다양성으로 인해 보다 정밀하고 신뢰성 높은 결과를 얻을 수 있다. 본 장에서는 이미지 분류에 사용되는 VLM 모델을 2.1, 2.2절에서, 객체 탐지에 사용되는 모델을 2.3, 2.4절에서 각각 살펴본다.

2.1 CLIP [1]

CLIP(Contrastive Language-Image Pre-training)[1]은 Vision-Language Model(VLM) 분야의 시초 모델이다. 전통적인 비전 모델은 사전에 정의된 고정 카테고리만을 인식하도록 학습되며, 새로운 개념을 도입하려면 추가적인 라벨링 데이터가 필요하다. 이에 반해 CLIP은 4억 개 이상의 이미지-텍스트 쌍과 함께 contrastive learning 방식으로 사전학습을 수행하여 추가적인 라벨링 데이터 없이도 새로운 카테고리를 인식할 수 있도록 한다. Fig. 1과 같이 CLIP은 image encoder와 text encoder로 구성되며, 모든 가능한 $N \times N$ 이미지-텍스트 조합 중에서 실제로 대응되는 쌍을 올바르게 식별하도록 학습한다. 즉, 이미지와 텍스트 임베딩 간의 실제 대응되는 쌍에 대한 유사도를 최대화하고, 대응되지 않는 쌍의 유사도를 최소화하도록 학습한다. 학습이 완료된 CLIP에 새로운 이미지와 텍스트를 입력하면, 해당 이미지에 대응되는 텍스트 프롬프트와의 유사도 점수를 기반으로 카테고리 라벨을 출력한다. 별도의 추가 학습 없이 zero-shot classification 능력을 발휘하므로, 라벨이 제한된 상황에서도 image classification을 수행할 수 있다. CLIP의 zero-shot 분류 메커니즘은 새로운 기뢰 형태나 다양한 기뢰 형태에 대한 추가적인 학습 없이도 효과적으로 대상 객체를 식별할 수 있어 기뢰 이미지 분류를 수행하는 데에 도움이 된다.

2.2 ALIGN[2]

기존의 시각 및 시각-언어 representation은 높은 라벨링 비용과 전문 지식이 요구되는 선별된 학습 데이터셋에 크게 의존한다. 이러한 정제된 데이터만을 선별하는 절차는 데이터셋 규모를 제한할 뿐만 아니라, 학습된 모델의 확장성을 저해한다는 한계를 갖는다. ALIGN(A Large-scale ImaGe and Noisy-text embedding)[2]은 이러한 문제점을 해결하기 위해

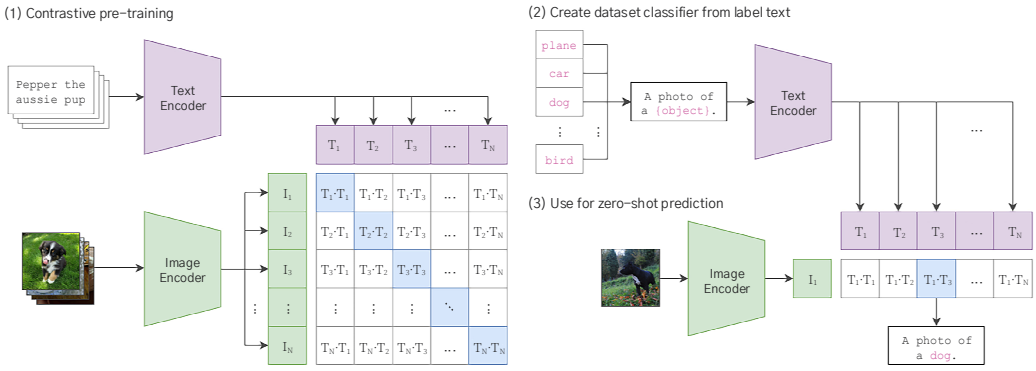

Fig. 1. Method of CLIP[1]

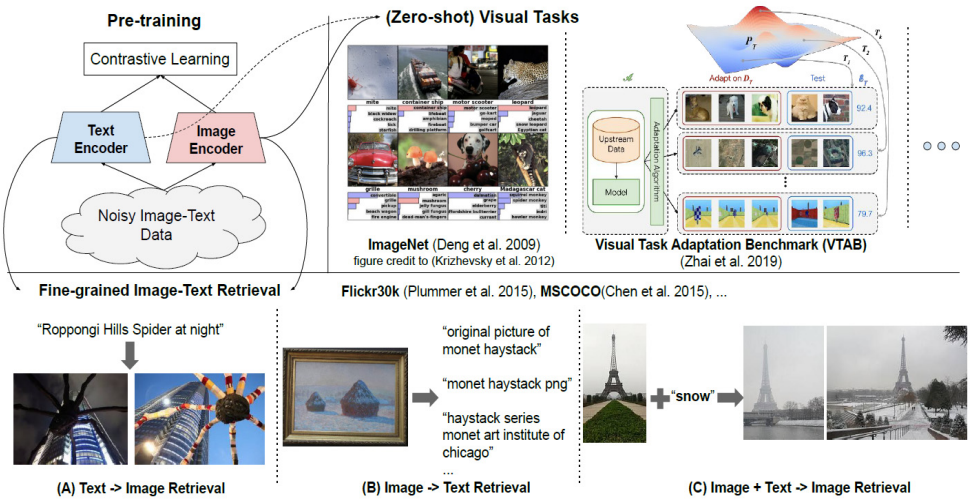

Fig. 2. Method of ALIGN[2]

필터링 및 후처리 단계 없이 수집된 10억 개의 이미지-대체 텍스트(alt-text) 쌍의 noisy 데이터셋을 활용한다. 비록 데이터에 다소 잡음이 존재하지만, 데이터셋 규모가 충분히 방대하기 때문에 잡음을 상쇄하고 최신 수준의 representation을 확보할 수 있다. Fig. 2와 같이 모델은 contrastive loss를 사용하여 이미지와 텍스트 쌍의 시각 및 언어 표현을 정렬하기 위해 학습하는 단순한 dual-encoder 구조로 구성한다. 이렇게 구성된 모델은 zero-shot image classification, image-text retrieval 등의 task를 수행한다. ALIGN 모델은 대규모의 데이터셋으로 사전학습되었기에, 처음 보는 기뢰 이미지를 분류 가능하다. 또한, 특정 기뢰를 찾기 위해, 기뢰의 형태를 서술하는 문장을 text prompt에 입력하여 찾고자 하는 기뢰 이미지를 image-text retrieval task를 통해 검색할 수 있다.

2.3 DetCLIP[3]

Open-world object detection을 수행하기 위한 기존 접근 중 하나인 GLIP[5]은 detection 데이터셋에 포함된 카테고리 명칭을 연속된 문장 형태로 연결하여 텍스트 프롬프트로 사용한다. 이는 텍스트 인코더의 입력 토큰 크기 제한으로 인해 많은 수의 카테고리를 다루거나 카테고리의 상세한 설명으로 확장하기 어렵다. 또한, 이는 연관 없는 카테고리 간의 비효율적인 상호작용을 초래한다. 이러한 한계를 극복하기 위해 DetCLIP[3]이 제안되었고, 모델의 구조는 Fig. 3와 같다. DetCLIP은 설계된 개념 사전으로부터 풍부한 사전 지식을 제공함으로써, open-world detection에 대한 병렬화된 시각-개념 사전학습 방식을 도입한다. 구체적으로, detection dataset, grounding dataset, image-text pair dataset과 같은 이질

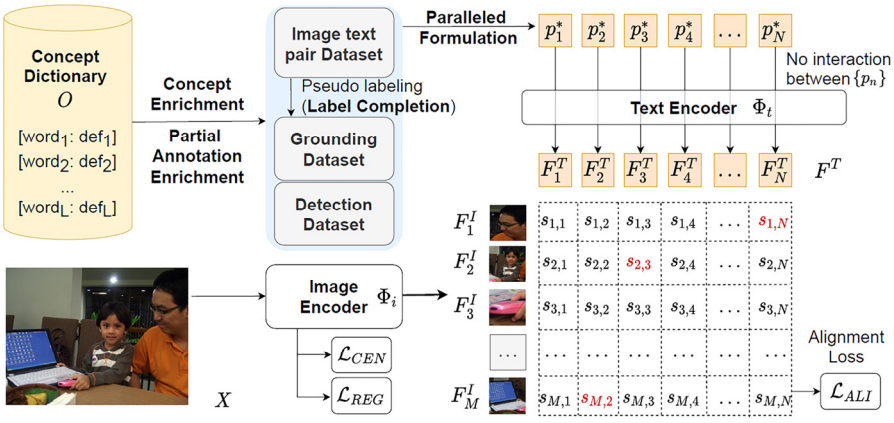

Fig. 3. Overall architecture of DetCLIP[3]

적인 데이터셋을 보다 효과적으로 활용하기 위해 개념을 분리해서 추출하는 병렬화된 개념 공식을 새롭게 설계하였다. 이 방식은 바운딩 박스에 대한 각 개념을 따로 떼어서 텍스트 인코더에 병렬로 넣는 입력 방식을 뜻한다. 이러한 병렬화된 개념 공식은 서로 관련되지 않은 카테고리들 간의 불필요한 상호작용을 방지한다. 또한, detection data, grounding data, image-text pair data를 병렬화된 방식으로 변환함으로써 서로 다른 형식을 가진 학습 데이터를 함께 사용하여 사전학습될 수 있으며, 이를 통해 객체 위치 탐지 능력과 추가 학습 없이 새로운 객체를 탐지할 수 있는 능력를 지원한다. 또한, DetCLIP은 다양한 온라인 자료와 기존 detection 데이터셋으로부터 구축된 개념 사전을 설계하여 각 개념에 대한 사전 지식을 제공한다.

2.4 YOLO-World[4]

YOLO는 객체 탐지 분야에서 많이 쓰이는 탐지 모델 중 하나다. 기존 YOLO 계열 모델은 사전에 정의되고 학습된 객체 카테고리에 의존하기 때문에, 새로운 클래스가 등장하거나 사전 정의되지 않은 객체가 존재하는 오픈 시나리오에서는 적용 가능성이 제한된다. 이러한 한계를 극복하고자 YOLO-World[4]가 제안되었다. YOLO-World는 vision-language 모델링과 대규모 이미지-텍스트 데이터셋을 이용한 사전학습을 통해, 기존 YOLO 구조에 open-vocabulary detection 능력을 부여한다. YOLO-World의 구조는 Fig. 4와 같다. 구체적으로는 Re-parameterizable Vision-Language Path Aggregation Network (RepVL-PAN)를 도입하여 시각적 정보와 언어적 정보 간의 상호작용을 효율적으로 통합한다. 또한, 시각적, 언어적 정보 사이의 상호작용을 용이하게 하기 위해 region-text contrastive loss를 제안한다. 이와 같은 시각적, 언어적 정보 간의 상호작용을 하는 설계 덕분에 YOLO-World는 zero-shot 방식으로 광범위한 객체를 탐지할 수 있다. 또한, YOLO-World는 정확도와 처리 속도 양 측면에서 우수함을 보이는 장점을 가지고 있기에, 수중 환경에서 기뢰 탐지 시스템을 구축하는 데 유용한 기반이 되며 실시간 처리가 요구되는 시스템에 효과적으로 적용할 수 있다.

3. 기뢰 탐지 및 분류에 VLM 적용 방안

2절에서 최근 연구에서 제안된 주요 VLM을 살펴보았다. 기존에 이미지 정보만으로는 한계가 있던 수중 기뢰 탐지·분류 문제에 VLM을 적용함으로써 해결 방안을 탐색한다. 이를 위해 기뢰 탐지·분류에 적합한 멀티모달 데이터셋을 구축하고, 구축된 데이터에 대해 목적에 부합하는 VLM을 선정하며, 선정된 VLM을 운용 환경에 배치 후 평가를 실시하는 순서로 적용 방안을 제시한다.

3.1 데이터셋 구축

수중 기뢰 탐지·분류를 위해서는 먼저 이미지-텍스트 쌍으로 이루어진 멀티모달 데이터셋을 구축하는 절차가 필요하다. 이때 데이터는 이미지와 텍스트를 쌍으로 이루어지도록 준비한다. 먼저, 이미지와 텍스트라벨을 포함하는 분류용 데이터셋을 만든다. 이미

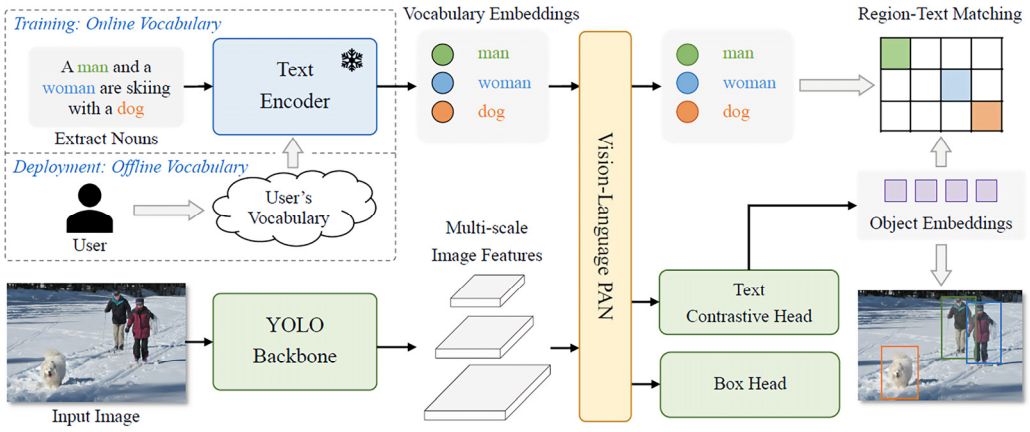

Fig. 4. Overall architecture of YOLO-World[4]

지는 수중 음향 센서로 촬영한 영상에서 얻으며, 동일한 기뢰를 다양한 방위와 거리, 조명 조건에서 촬영해 환경 변화에도 강인한 데이터를 확보한다. 텍스트는 적용하고자 하는 탐지·분류 모델의 특성에 따라, 기뢰 단어나 기뢰의 형태, 재질, 상태 등을 기술한 문장으로 작성한다. 이렇게 구성된 이미지-텍스트 쌍은 VLM이 추가 학습하고, 분류 성능을 평가하는 기본 자료가 된다.

분류용 데이터가 완성된 뒤에는 탐지용 데이터셋을 추가로 구축한다. 탐지용 이미지는 분류용 이미지 데이터셋에 기뢰가 차지하는 영역을 직사각형 바운딩 박스로 어노테이션한 후, 각각의 박스에 해당하는 라벨을 부여한다. 최종적으로는 이미지, 바운딩 박스, 텍스트 라벨을 통합하여 학습 및 평가 단계에서 사용할 수 있도록 한다.

3.2 VLM 선정 및 성능 평가

VLM을 기뢰 탐지·분류에 적용하기 위해, 목적에 따라 적절한 모델을 선정하고 이를 실제 운용 환경에 배치한 뒤 성능을 평가하는 일련의 절차를 기술한다. 우선, 기뢰 이미지 분류와 이미지 내 기뢰 탐지라는 두 가지 핵심 문제를 구분한다. 기뢰 이미지 분류의 경우, CLIP[1] 또는 ALIGN[2]과 같은 분류 VLM을 사용한다. 반면, 이미지 내에서 기뢰를 탐지해야 하는 경우에는, DetCLIP[3] 또는 YOLO-World[4]와 같은 객체 탐지 VLM을 사용할 수 있다.

선정된 모델을 실제 운용 환경에 적용하기 위해 edge-device 혹은 함정 내 서버에 모델을 배치한다. 이를 위해 먼저 하드웨어 사양을 정의하고, 딥러닝 학습을 위한 소프트웨어를 설치한다. 추가 학습 및 추론 단계에서는 모델에 따라 3.1절에서 구축한 이미지-텍스트 쌍을 활용한다. zero-shot 이미지 분류에서는 학습 없이 기뢰 이미지와 텍스트 프롬프트를 입력으로 제공하면, 이미지에 대해 예측한 라벨이 출력된다. zero-shot 객체 탐지에서는 이미지와 텍스트 프롬프트를 입력으로 제공하면, bounding box 좌표와 confidence score가 산출된다.

이와 같이 VLM은 재학습 없이 적용 가능한 zero-shot 인식을 제공하며, 운용 환경에서 새롭게 등장하는 기뢰 유형에 대해서도 인식 범위를 확장할 수 있어 open-world 탐지 시나리오에 적합하다. 또한, 필요 시 소량의 데이터에 기반한 추가 학습을 통해 특정 기뢰 유형이나 운용 환경에 대한 인식 성능을 점진적으로 향상시킬 수 있다는 장점을 지닌다. 특히, 다양한 전술과 작전 환경에 따라 목표하는 기뢰를 탐지하는 데에 적용할 수 있다.

성능 평가 지표는 분류 문제에 대해 accuracy를, 탐지 문제에 대해서는 mAP(mean Average Precision)를 사용하여 분류 및 탐지 정확도를 평가한다. 또한 실시간 운용을 위한 초당 처리 가능한 프레임 수(fps)와 메모리 사용량을 측정하여, 실시간성과 경량성을 가지는지 검증한다.

4. 결론

본 연구는 최신 비전-언어 모델(VLM)의 최신 연구 동향을 분석하고, 이를 수중 기뢰 탐지·분류에 적용할 수 있는 방법론을 제시하였다. 대규모 이미지-텍스트 쌍으로 사전 학습된 CLIP[1]·ALIGN[2]은 이미지 분류 능력을, DetCLIP[3]·YOLO-World[4]는 텍스트 기반 객체 검출 성능을 각각 제공한다는 점을 확인하였다. 이에 따라 기뢰 분류에는 CLIP·ALIGN을, 기뢰 탐지에는 DetCLIP·YOLO-World를 선택하고, edge-device 혹은 함정 내 서버에 배포할 수 있다. 이러한 VLM 적용 방식은 새로운 기뢰 유형, 수중 환경 변화, 다변화된 작전 환경에 대해 적응력있는 기뢰 탐지 시스템으로 활용될 것으로 기대한다. 앞으로 기뢰 탐지·분류를 위한 VLM 활용 방안에 대한 지속적인 연구가 필요하며, 더 나아가 데이터베이스 확장, 프롬프트 튜닝기법, 경량화 기법과 같은 기뢰 탐지·분류 시스템을 위한 심도 있는 연구가 필요하다.

참고문헌

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, & Ilya Sutskever, 'Learning Transferable Visual Models from Natural Language Supervision,' in proceedings of the 38th International Conference on Machine Learning, 2021, pp. 8748-8763.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, & Tom Duerig, 'Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision,' in

proceedings of the 38th International Conference on Machine Learning, 2021, pp. 4904-4916.
[3] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, & Hang, Xu, 'Detclip: Dictionary-Enriched Visual-Concept Paralleled Pre-Training for Open-World Detection,' Advances in Neural Information Processing Systems 35, 2022, pp. 9125-9138.
[4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, & Ying Shan, 'Yolo-World: Real-Time Open-Vocabulary Object Detection,' in proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901-16911.
[5] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, & Jianfeng Gao, 'Grounded Language-Image Pre-Training,' in proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10965-10975.