객체 마스크 단위

전역 특징 정렬을 활용한

위장 객체 탐지 모델 성능 향상 방법

A Method for Enhancing

Camouflaged Object Detection Performance

via Mask-level Global Feature Alignment

서강대학교

컴퓨터공학과

홍 승 훈

요 약

실시간 위장 객체 탐지(Real-time Camouflaged Object Detection) 모델은 30FPS 이상의 속도를 가지면서, 위장 객체에 특화된 탐지 능력을 갖춘 모델을 의미한다. 이는 위장된 객체에 대한 즉각적인 반응이 필수적인 국방, 감시 시스템에서 핵심적인 기술이다. 하지만 일반 객체와 달리 위장된 객체는 배경과 구분이 모호하여 정확한 탐지를 위해 추가적인 모델 설계가 필요하고, 실시간 처리를 위해서는 연산 효율성까지 요구되어 상충되는 문제가 존재한다. 본 논문에서는 실시간 객체 탐지 모델의 위장 객체에 대한 정확도를 향상시키기 위해, 실시간 객체 탐지 모델인 D-FINE 모델을 기반으로 '객체 마스크 단위 전역 특징 정렬을 활용한 효과적인 인코더 학습 방법'을 새롭게 제안한다. 제안하는 방법은 학습 시, 객체 영역을 정확히 추출하기 위해 각 스케일의 특징 맵 내 정답 객체 영역에서 채널 방향 L2 정규화를 한 값이 가장 큰 상위 K개(Top-K)의 핵심 요소를 동적으로 선택한다. 이후 이로부터 각 스케일에서의 객체를 대표하는 '프로토타입 벡터'를 생성하고, 분포 손실 함수를 통해 하위 스케일(local, medium)의 프로토타입을 상위 글로벌 스케일의 프로토타입과 KL divergence로 정렬시킨다. 제안하는 방법은 추론 시에는 추가 연산 비용 없이, 오직 학습 과정에서만 다중 스케일 특징 맵의 의미론적 일관성을 강화한다. 제안 방법을 활용해 학습한 모델은 군사용 위장 객체 탐지 데이터셋인 MHCD2022에서 기존 모델 대비 우수한 성능을 나타낸다 . 이는 제안하는 방법이 객체 탐지 모델의 위장 객체에 대한 다중 스케일 특징 표현을 효과적으로 강화하여, 위장 객체의 탐지 정밀도를 크게 높일 수 있음을 보여준다.

주제어: 실시간 위장 객체 탐지, 다중 스케일 특징, 특징 정렬, 프로토타입

Abstract

Real-time Camouflaged Object Detection models are defined as those capable of operating at speeds exceeding 30 FPS while possessing specialized detection capabilities for camouflaged objects. This technology is critical in defense and surveillance systems where immediate reactions to concealed targets are essential. However, unlike general objects, camouflaged objects have ambiguous distinctions from their background, requiring additional model design for accurate detection. This creates a conflict with the computational efficiency required for real-time processing. In this paper, to improve the accuracy of camouflaged object detection based on the real-time object detection model D-FINE, we propose a novel "effective encoder training method utilizing object mask-level global feature alignment." During training, to accurately extract object regions, the proposed method dynamically selects the Top-K key elements with the largest channel-wise L2 normalized values within the ground truth object area in the feature maps of each scale. Subsequently, "prototype vectors" representing the object at each scale are generated, and the prototypes of lower scales (local, medium) are aligned with the prototype of the upper global scale using KL divergence through a distribution loss function. This method strengthens the semantic consistency of multi-scale feature maps solely during the training process, incurring no additional computational cost during inference. The model trained using the proposed method shows superior performance compared to existing models on the MHCD2022 military camouflaged object detection dataset. This demonstrates that the proposed method effectively enhances the multi-scale feature representation for camouflaged objects, significantly improving detection precision.

Topics : Real-time Camouflaged Object Detection, Multi-scale Feature, Feature Alignment, Prototype

목 차

그림 목차

표 목차

1. 서론

객체 탐지 분야에서 실시간 객체 탐지(Real-time Object Detection)[1] 기술은 자율주행, 감시 시스템 등 즉각적인 반응을 요구하는 실시간 응용 분야에서 매우 중요하다. 또한, 국방과 같은 영역에서는 실시간으로 위장한 객체를 성공적으로 탐지하는 것이 작전의 성패에 큰 영향을 주기도 한다. 이와 같은 실무적인 문제와는 달리, 학술적인 분야에서는 실시간 객체 탐지와 위장 객체 탐지(Camouflaged Object Detection)[2]가 별개의 영역으로 발전되어 왔다. 실시간 객체 탐지는 MSCOCO 데이터셋[3]과 같이 일상적인 객체로 이루어진 Bounding Box(Bbox) 형태의 벤치마크 데이터셋을 활용하고, 위장 객체 탐지는 COD10K[4] 데이터셋을 활용하여 실시간 탐지보다는 정확도에 초점을 맞춘 Segmentation을 위주로 발전되어 왔다. 따라서, 국방과 보안에 관련된 실무적인 목적에 맞게 실시간성과 정확도 모두를 고려한 실시간 위장 객체 탐지 모델에 대한 고민이 필요하다.

일반적인 Bbox 형태의 객체 탐지가 아닌 Segmentation을 활용한 객체 탐지는 모든 픽셀의 클래스를 예측해야 하기 때문에 아무리 효율적인 시스템을 구축한다고 하더라도, 근본적인 연산량의 차이로 인해 실시간 측면에서 Bbox 형태의 실시간 객체 탐지 모델을 따라갈 수 없다. 따라서, 위장 객체 탐지 모델의 실시간성을 확보하기 위해서는 Bbox 형태의 결과를 내는 실시간 객체 탐지 모델을 선택하고, 위장 객체를 더 효과적으로 탐지할 수 있도록 모델을 재설계하는 것이 필요하다. 즉, 일반 객체와 위장 객체에 대한 근본적인 차이를 이해하고 모델을 재설계 해야 한다. 많은 위장 객체 탐지에 관한 논문들에서는 가장 두드러지는 차이는 배경과 전경의 차이가 불분명하여 경계를 찾기가 힘들다는 점이라고 지적한다[5]. 이로 인해 객체 탐지 분야에서 주로 활용되는 FPN과 같은 구조에서는 해상도가 다른 특징 맵을 의미론적으로 융합하지 못한다[6]. 이와 같은 문제를 해결하기 위해 객

체 탐지 영역에서는 프로토타입(prototype)을 활용하여 각 객체에 대한 의미론적 접근에 대한 연구가 활발하게 진행되고 있다. 객체 클래스에 대한 프로토타입을 별도의 파라미터로 저장하거나, 클러스터링을 통해 유사도를 측정하는 방식이 대표적이다[7][8].

본 논문에서는 이러한 다중 스케일 특징 간의 의미론적 불일치 문제를 해결하기 위해, 실시간 탐지가 가능한 D-FINE[9] 모델을 기반으로 '객체 마스크 단위 전역 특징 정렬을 활용한 효과적인 인코더 학습 방법'을 새롭게 제안한다. 제안하는 방법은 인코더를 통과한 다중 스케일 특징 맵 상의 정답 객체 영역 내에서 피처의 정규화된 값을 기준으로 가장 의미론적으로 중요하다고 판단되는 상위 K개의 핵심 요소(Top-K components)를 동적으로 선택한다. 이후, 각각의 스케일에서 채널별 평균과 Power iteration을 통한 주성분을 계산하고, 이를 각 스케일을 대표하는 '프로토타입 벡터'로 지정한다. 기준이 되는 상위 스케일의 프로토타입과 나머지 스케일의 프로토타입의 KL divergence를 계산하여 하위 스케일의 프로토타입이 상위 스케일의 프로토타입과 정렬되도록 한다. 이 방법의 핵심은 추론 시에는 어떠한 추가 연산 비용도 발생시키지 않고, 오직 학습 과정에 사용되는 손실을 추가하여 모델의 특징 표현 능력을 극대화하는 것이다. 이 방법을 활용하면 실시간 객체 탐지 모델의 실시간성을 잃지 않으면서도, 위장 객체 탐지에 강건한 모델을 만들 수 있다.

본 논문의 실험을 위한 학습 및 평가는 군사용 위장 객체 탐지 데이터셋인 MHCD2022[10]를 활용했다. 제안 모델은 기존 모델 대비 모든 지표에서 성능이 향상되었다. mAP@0.5 지표는 77.94%로 2.67%p 상승했으며, mAP@0.75 지표는 61.49%로 5.53%p, mAP@[0.5:.95] 지표는 54.95%로 3.24%p 상승했다. 이는 제안방법이 실시간성을 유지하면서도, 다중 스케일 특징의 불일치 문제를 효과적으로 해소하여 위장객체탐지의 정밀도를 크게 향상시킬 수 있음을 입증한다.

본 논문의 구성은 서론을 제외한 4장으로 이루어져 있으며 제2장에서는 실시간 객체 탐지, 위장 객체 탐지 및 특징 정렬에 대한 관련 연구를 살펴보고 기존 연구 적용 시의 문제점을 분석한다. 제3장에서는 제안하는 전역 특징 정렬을 활용한 인코더 학습 방법을 포함한 전체 시스템 설계를 상세히 다루고, 제4장에서는 기준 모델과의 정량적, 정성적 비교를 통해 제안 시스템의 성능을 분석한다. 제5장에서는 본 논문에서 제안한 방법의 결론과 향후 연구 방향을 제시한다.

2. 연구 배경

본 장에서는 위장객체탐지의 문제 정의와 실시간 처리가 요구되는 분야에서의 문제점을 정의한다. 이어서, 본 연구의 기반이 되는 실시간 객체탐지 모델과 기존 위장객체탐지 연구를 비교 분석하고, 객체탐지 분야에서 사용되는 특징 정렬 방법들에 대해 고찰하며 본 논문의 접근 방향을 제시한다.

2.1 실시간 위장 객체 탐지 문제 정의

위장 객체 탐지 기술은 객체가 주변 배경과 시각적으로 유사하여 구분이 어려운 환경에서 객체를 배경과 구분하는 기술[2]을 말한다. 이러한 기술은 특히 정찰 드론, 감시 시스템, 표적 식별 등 신속하고 정확한 인식이 요구되는 군사 분야에서 핵심적인 중요성을 갖는다.

하지만 기존의 위장 객체 탐지 연구는 주로 픽셀 단위의 정밀한 마스크를 예측하는 Segmentation 작업에 집중되어 왔다. Segmentation은 이미지의 픽셀 단위 클래스 예측으로 객체의 정확한 윤곽을 파악할 수 있다는 장점이 있으나, 모든 픽셀에 대해 클래스를 예측해야 하므로 Bbox 대비 막대한 연산량을 요구한다. 이는 추론 지연을 발생시켜 30FPS 이상의 실시간 탐지가 필수적인 군사적 분야에서는 부적합하다. 따라서 본 연구에서는 실시간성이 보장되는 Bbox 기반의 객체 탐지 방식으로 위장 객체 탐지 문제에 접근하고자 한다. 그러나 단순히 Bbox 형태의 일반 객체 탐지 모델을 위장 객체 데이터로 학습하는 것은 한계가 있다. 위장 객체는 본질적으로 배경과 객체의 특징 표현(Feature Representation)이 매우 유사하여, 모델이 객체와 배경을 구분하는 데 큰 어려움을 겪는다. 일반 탐지 모델은 이러한 이유로 인해 전경을 배경으로 오인하거나, 객체의 정확한 위치를 특정하지 못하는 경향이 있다.

결론적으로, 본 연구가 해결하고자 하는 핵심 문제는 "실시간 추론 속도를 유지하면서도, 배경과 혼재된 위장 객체의 약한 특징을 효과적으로 학습하여 탐지 정확도를 극대화하는 Bbox 기반 탐지 모델을 설계"하는 것이다. 이를 위해 본 논문에서는 실시간 객체 탐지 모델인 D-FINE[9]을 기반으로, 위장 객체의 의미론적 특징(Semantic Feature)을 정렬하는 새로운 학습 방법을 제안한다.

2.2 관련 연구

본 절에서는 실시간 객체 탐지 모델과 위장 객체 탐지 분야의 연구 흐름을 살펴보고, 일반 객체와 다른 위장 객체 탐지의 특수한 문제점을 고찰한다. 또한, 객체 탐지 분야에서 주목받고 있는 특징 정렬(Feature Alignment) 방법[11]과 프로토타입 정렬(Prototype Alignment)[7] 연구를 함께 검토한다. 이를 통해 본 논문에서 제안하는 전역 특징 정렬을 활용한 위장 객체 탐지 모델 학습방법의 필요성과 이론적 근거를 제시한다.

2.2.1 실시간 객체탐지 모델

실시간 객체 탐지 모델은 제한된 연산 자원 하에서 30FPS 이상의 추론 속도와 높은 검출 정확도를 동시에 달성하는 것을 목표로 한다. 이러한 모델은 자율주행, 영상 감시, 로봇 비전 등 시간 지연(latency)이 치명적인 응용 환경에서 특히 중요하게 다루어진다.

초기의 실시간 객체탐지 연구는 YOLO(You Only Look Once)[1]와 SSD(Single Shot MultiBox Detector)[12]을 기점으로 시작되었다. 이들 모델은 별도의 후보 영역 생성(region proposal) 단계를 제거하고, 하나의 신경망을 통해 Bbox와 클래스를 동시에 예측하는 구조를 제안하였다. 이와 같은 구조를 one-

stage 방식으로 부르며, 기존의 two-stage 방식보다 훨씬 빠른 속도를 확보하였으나, 정밀도(accuracy)가 다소 낮고 작은 객체 검출에 취약한 한계를 보였다.

이후, EfficientDet[13]과 YOLOv7부터 v12까지 경량 백본(lightweight backbone)과 Feature Pyramid Network(FPN)[11]또는 Path Aggregation Network(PAN)[14]을 결합하여, 다양한 크기의 객체에 대응할 수 있도록 하였다. FPN과 PAN은 각 계층의 의미적 특징(semantic feature)과 세부적 특징(detail feature)을 정렬(align)하여, 스케일 간 일관성(scale consistency)을 보장한다. 이러한 구조는 객체 탐지 모델에서 필수적인 요소로 자리 잡았다.

최근에는 Transformer 기반 구조가 실시간 객체 탐지 분야로 활발히 확장되고 있다. Transformer 기반 객체 탐지 모델인 DETR(Detection Transformer)[15]은 NMS나 앵커 박스(Anchor Box)에 의존하지 않고 End-To-End 방식으로 검출을 수행하는 모델이다. 하지만, 높은 지연 시간(high latency)와 연산량으로 실시간 모델로 활용하기 부적합하여 이를 실시간 모델로 개량한 RT-DETR[16]과 LW-DETR[17] 모델 등이 등장했다. 특히, 최근 제안된 D-FINE(Dynamic Feature Interaction Network for Efficient Detection)[9] 모델은 실시간 환경에서 정밀한 검출을 달성하기 위한 구조로 주목받고 있다. D-FINE은 기존 DETR 계열 모델의 coarse-to-fine refinement 구조를 개선하여, 세밀한 특징 정제(Fine-grained Feature Refinement)와 전역 정렬(Global Alignment)을 동시에 수행한다. 이를 통해 스케일별 특징 불일치(misalignment) 문제를 해결하고, 작은 객체나 복잡한 배경 내 객체의 위치 예측 정확도를 높였다. 또한 D-FINE은 모델 전체가 효율적으로 경량화되어, GPU 및 엣지 디바이스 환경에서도 안정적인 실시간 성능을 보장한다.

이러한 흐름은 단순히 추론 속도의 향상뿐 아니라, 스케일 간 정렬(Scale Alignment)과 의미적 정렬(Semantic Alignment)을 실시간성 하에서도 유지하는 방향으로 발전하고 있다. 본 논문에서는 이러한 정렬 개념을 위장 객체에 대한 개념까지 확장하여, 위장 객체 탐지 상황에서 발생하는 스케일 간 특징 맵 불일치 문제를 해결하기 위한 학습 방법을 제안한다.

2.2.2 위장 객체 탐지 모델

위장 객체 탐지는 객체가 주변 배경과 시각적으로 유사하여, 인간의 시각체계나 일반적인 탐지 모델이 인식하기 어려운 환경에서 목표 대상을 찾아내는 기술이다. COD의 핵심 문제는 객체와 배경 간의 시각적 유사성으로 인해 특징 간의 융합이 어렵다는 점이다[18]. 즉, 위장 객체는 동일한 특징 공간에서 배경 특징와 혼재되기 쉬우며, 스케일별 특징 분포의 일관성이 약화되는 경향을 보인다.

기존 COD 연구들은 COD10K[4] 데이터셋을 활용해 주로 Segmentation 기반으로 접근하여, 객체의 경계를 정밀하게 복원하는 데 중점을 두었다. 예를 들어, SINet[4]은 Context Aggregation 모듈을 통해 지역/전역 특징을 통합하였고, ZoomNet[19]과 CSGnet[20]은 Progressive Zoom Mechanism을 도입하여 다중 스케일의 특징을 재정렬함으로써 위장 객체의 경계 정보를 강화하였다. 그러나 ZoomNet에서도 언급되었듯이, 위장 객체의 스케일별 표현은 여전히 정렬 불안정성(scale-wise instability)을 보이며, 작은 객체나 저대비(low-contrast) 영역에서는 성능 저하가 발생한다.

하지만 이러한 Segmentation 기반 접근은 실시간성 측면에서 한계를 가진다. 마스크 단위의 픽셀 예측은 연산량이 많아 지연(latency)이 발생하며, 국방과 감시 등의 실시간 응용에서는 즉각적인 반응 속도를 보장하기 어렵다. 따라서 최근

연구들은 COD 문제를 Segmentation(mask) 기반이 아닌, 검출(bounding box, bbox) 기반 문제로 재정의하려는 시도를 보이고 있다[10]. 이러한 방식은 객체의 픽셀 단위 복원보다는 위치(localization)에 집중함으로써, 계산 효율성과 속도 모두를 확보할 수 있다.

이에 따라 군사적 목적의 응용 환경에서는 마스크 단위의 복잡한 경계 정보보다는 객체의 위치 정보(bounding box)를 빠르고 안정적으로 파악하는 것이 더 효율적이다. 실시간 표적 탐지나 영상 감시 시스템에서는 짧은 시간 내에 위장된 객체의 존재 여부와 위치를 확인하는 것이 중요하기 때문이다. 이러한 이유로, 군사적 상황에서는 bbox 기반 접근이 실시간 위장객체탐지에 보다 적합한 방식으로 간주된다.

2.2.3 특징 정렬에 관한 연구

객체 탐지 및 위장 객체 탐지의 성능을 결정짓는 핵심 요인 중 하나는 특징 정렬(Feature Alignment)이다. 특징 정렬은 서로 다른 스케일 혹은 네트워크 계층에서 추출된 특징들이 의미적으로 동일한 객체를 일관성 있게 표현하도록 학습하는 과정으로, 특히 위장 객체처럼 시각적으로 배경과 혼재된 대상에서는 그 중요성이 더욱 강조된다.

초기 특징 정렬 연구는 FPN 구조[11]에서 출발하였다. FPN은 상위 레벨의 의미적 특징과 하위 레벨의 세부적 특징을 결합하여 다중 스케일 간 특징 정렬을 수행한다. 이 방식은 객체의 크기나 위치에 무관하게 안정적인 검출이 가능하도록 하며, 이후 대부분의 탐지 네트워크의 기본 구조로 채택되었다. 그러나 FPN은 위치적 정렬(spatial alignment)에는 효과적이지만, 의미적 정렬(semantic alignment)까지 보장하지는 못한다는 한계가 있다.

이후 DETR[15] 계열의 연구에서는 품질 정렬(Quality Alignment) 개념이 새롭게 등장하였다. 대표적으로 Align-DETR[21]은 기존 DETR이 가지는 분류(Classification)와 회귀(Regression) 간의 비정렬 문제를 지적하였다. 동일한 특징이 위치 예측에는 적합하지만, 범주 예측에는 부정확할 수 있다는 점에서, Align-DETR은 두 출력을 정렬시키기 위한 Aligned Matching Loss를 제안하였다. 이 접근은 단순히 공간적인 정렬뿐만 아니라, 특징 표현 간의 의미적 일관성(semantic consistency)을 보장하기 위한 정렬 학습의 중요성을 보여준다.

한편, 위장 객체 탐지와 같이 객체와 배경 간 경계가 불명확한 문제에서는, 특징 정렬을 클래스 단위로 확장한 프로토타입 정렬(Prototype Alignment) 접근이 연구되고 있다. PANet[7]와 Context Prototype-Aware Learning[8]에서는 클래스별 프로토타입 벡터를 구축하여, 각 특징이 해당 프로토타입에 수렴하도록 학습함으로써 특징 분포 간의 일관성을 향상시켰다. 이러한 방식은 특징 공간(feature space)에서의 전역적인 의미 정렬(global semantic alignment)을 유도하고, 클래스 간 구분성을 강화하는 효과가 있다.

더 나아가, 최근에는 특징와 프로토타입 간의 정렬을 최적수송(Optimal Transport, OT)기반으로 공식화한 연구들이 제안되고 있다. 대표적으로 POT[22]은 특징 분포와 프로토타입 분포 간의 거리를 최소화하는 최적수송 함수를 통해, 특징 간의 대응관계를 정량적으로 모델링하였다. POT은 각 특징이 어떤 프로토타입에 정렬되어야 하는지를 확률적으로 계산하며, 학습 과정에서 글로벌 특징 정렬(global feature alignment)을 자연스럽게 유도한다. 이러한 접근은 복잡한 배경 내 객체 분포가 불균형한 위장객체탐지 문제에서도, 특징 간 불일치를 완화하는 효과가 있다.

본 논문에서는 이러한 연구 흐름을 바탕으로, 위장 객체 탐지에서의 전역 특징 정렬(Global Feature Alignment)개념을 제안한다. 특히 D-FINE[9] 모델의 전역 특징 맵을 정렬 대상으로 설정하고, 스케일별 특징 간의 분포 일관성을 유지한 상태에서 지역 특징이 집중해야 할 영역을 학습함으로써, 위장된 객체의 시각적 혼재 문제를 해결하고자 한다.

2.3 기존 연구 적용 시의 문제점

앞서 2.2절에서 살펴보았듯이, 실시간 객체 탐지(2.2.1절)와 위장 객체 탐지(2.2.2절) 연구는 각기 다른 목표를 가지고 발전해왔다. 실시간 객체 탐지는 '속도'와 '효율'을 최우선으로 하며 경량화된 Bbox 기반 구조를 채택하는 반면, 위장객체탐지는 '정밀도'와 '경계 구분'을 위해 Segmentation 기반으로 연구가 집중되어 왔다.

이러한 연구 방향의 차이는 실시간 위장객체탐지, 특히 군사적 목적의 응용(MHCD2022[10] 데이터셋)에 기존 연구를 직접 적용할 때 명확한 한계점을 야기한다.

첫째, 기존 위장 객체 탐지 모델의 계산 복잡성 문제이다. 2.2.2절에서 언급된 SINet[4], ZoomNet[19] 등 대다수의 COD 모델은 픽셀 단위의 마스크(mask)를 예측하는 Segmentation 방식을 채택한다. 이 방식은 $H \times W$해상도의 특징 맵에 대해, 모든 픽셀($H \times W$개)마다 위장 객체인지 배경인지를 분류하는 조밀한(dense) 연산을 수행한다. 최종 예측을 위해 ($H \times W \times C$(클래스 수)) 차원의 출력을 생성해야 하므로 막대한 계산량을 필요로 한다. 반면, D-FINE[9]과 같은 DETR[15] 계열의 Bbox 모델은 N개의 쿼리(N=300~900)에 대해서만 예측을 수행한다. 즉, 최종 출력 헤드는 분류(Class Head)와 회귀(Bbox Head) 두 부분으로 나뉘며, 연

산량은 N × C(분류) 및 N × 4(Bbox 좌표)에 비례한다. N은 H × W보다 현저히 작으므로, 이 희소한(sparse) 예측 방식은 높은 추론 지연 시간(latency)을 근본적으로 해결하며 군사적 환경의 실시간성을 보장할 수 있다.

둘째, 기존 데이터셋과 군사적 응용 목적 간의 불일치이다. 기존 COD 연구를 주도한 COD10K[4]와 같은 벤치마크는 자연 속 동물이나 곤충을 포함한다. 이 데이터셋의 Bbox는 종종 객체보다 훨씬 넓은 영역을 포함하여, Bbox 대비 실제 객체가 차지하는 픽셀 비율이 매우 낮다. 즉, Bbox 내에 과도한 양의 배경 정보가 포함된다. 이러한 특성은 Bbox 전체의 특징을 이용해 위치를 회귀해야 하는 Bbox 탐지 모델의 학습을 방해하며, 배경 노이즈로 인해 성능이 저하된다. 따라서 COD10K와 같은 데이터셋은 Bbox 탐지에 비효율적이다. 하지만 MHCD2022[10] 데이터셋에 포함된 탱크, 전투기, 함정, 사람 등은 상대적으로 명확한 형태를 가지며 Bbox 내 객체 비율이 높다. 이는 Bbox 탐지 모델이 배경 노이즈의 영향을 덜 받고 객체 특징에 집중할 수 있음을 의미하며, 군사적 목적의 실시간 탐지에는 Bbox 기반 접근이 더 효과적일 수 있음을 시사한다.

결론적으로, 군사적 응용을 위한 실시간 위장객체탐지 시스템을 구축하기 위해서는 기존의 Segmentation 중심 접근에서 벗어날 필요가 있다. 연산량이 적은 Bbox 형태의 객체탐지 모델을 기본 구조로 채택하되, 위장 객체 고유의 문제인 '배경과의 시각적 유사성'과 '특징 불일치' 문제를 효과적으로 해결할 수 있는 새로운 학습 방법이 요구된다. 따라서, 본 논문은 실시간 bbox 탐지기의 효율성을 유지하면서 위장객체탐지 성능을 극대화하는 특징 정렬 방법을 제안하고자 한다.

3. 객체 마스크 단위 전역 특징 정렬을 활용한 위장 객체 탐지 모델 성능 향상 방법

본 장에서는 본 연구에서 제안하는 '객체 단위 전역 특징 정렬(Object-level Global Feature Alignment)' 방법의 설계와 구성에 대해 설명한다. 제안하는 방법은 Bbox 내에서 객체의 중요한 정보만을 선택적으로 추출(3.2절)하고, 이를 안정적인 전역 특징 벡터(Global Feature Vector) 기준으로 정렬(3.3절)함으로써, 다중 스케일 특징 맵 전체의 의미론적 일관성을 강화한다.

3.1 전체 시스템 구조

본 절에서는 객체 단위 전역 특징 정렬을 통한 객체 탐지 위장 객체 탐지 모델 성능 향상 방법에 대한 전체적인 구조를 설명한다. 연구에서 제안하는 모델의 전체 시스템 구조는 <그림 1>과 같다.

먼저, 상단의 기존 Baseline은 D-FINE[9] 모델의 순전파 과정을 설명한다. 순전파 과정에서 모델은 CNN 기반의 백본 네트워크를 통해 다중 스케일의 원본 특징 맵(local, medium, global)을 추출하고, 1x1 합성곱과 트랜스포머 인코더를 거쳐 평탄화한 뒤 트랜스포머 디코더의 입력으로 활용된다. 이후 트랜스포머 디코더를 통과 한 후 FFN을 통해 4개의 Bbox 정보를 생성하고, Ground Truth(GT)과의 비교를 통해 Bbox Loss를 계산하여 전체 네트워크를 업데이트한다.

이 때, 본 논문에서 제안하는 시스템을 통해 계산된 Distribution Loss는 Bbox Loss에 더하여 역전파 과정에서 활용된다. 트랜스포머 인코더를 통과한 평탄화 직전의 다중 스케일 인코딩된 특징 벡터(local, medium, global)를 활용하여 관심 영역 추출기와 프로토타입 정렬 모듈을 거쳐 각 스케일별 특징 맵의 프로토타입을 비교해 Distribution Loss를 계산한다. 이는 local, medium 스케일의 특징 벡터를 인코딩하는 1x1 합성곱 필터를 효과적으로 학습시킨다.

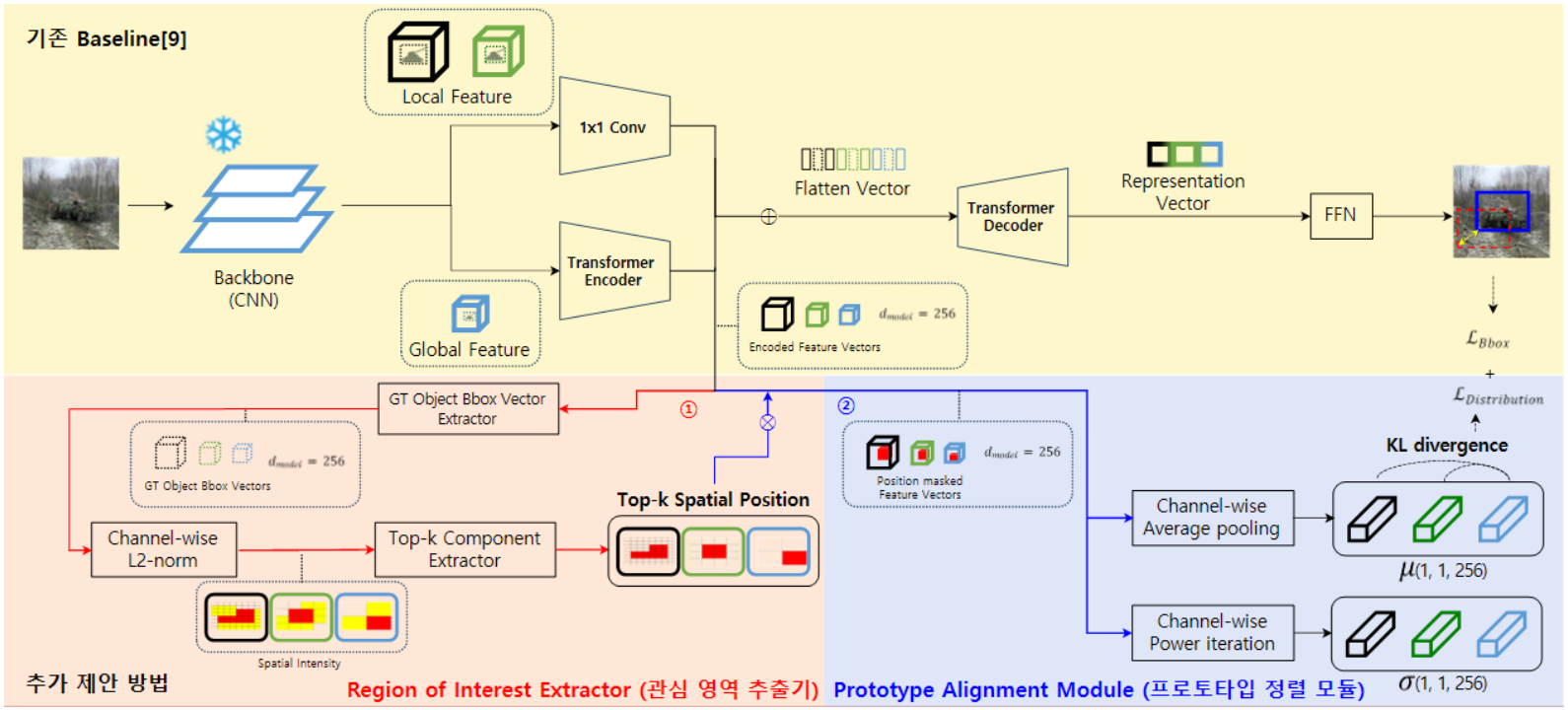
기존 Baseline[9]
Local Feature
1x1 Conv
Backbone
(CNN)
Transformer
Encoder
Global Feature
Flatten Vector
Transformer
Decoder
Representation
Vector
FFN
$\mathcal{L}_{Bbox}$
+
$\mathcal{L}_{Distribution}$
KL divergence
Encoded Feature Vectors
$d_{model} = 256$
GT Object Bbox Vector
Extractor
GT Object Bbox Vectors
Channel-wise
L2-norm
Top-k Component
Extractor
Spatial Intensity
Top-k Spatial Position
Position masked
Feature Vectors
Channel-wise
Average pooling
Channel-wise
Power iteration
$\mu(1, 1, 256)$
$\sigma(1, 1, 256)$
추가 제안 방법
Region of Interest Extractor (관심 영역 추출기)
Prototype Alignment Module (프로토타입 정렬 모듈)

그림 1 전체 시스템 구조

3.2 객체 마스크 단위 전역 특징 정렬을 위한 세부 시스템 설계 방법

3.2.1 관심 영역 추출기 설계

관심 영역 추출기는 다중 스케일 특징 맵에서 객체의 핵심 정보만을 정밀하게 선별하기 위해 설계된 모듈이다. 이 모듈은 Bounding Box 내의 모든 정보를 사용하는 대신, 특징 활성도(Feature Activation)가 높은 상위 픽셀만을 추출하여 배경 잡음을 제거하고 특징 정렬의 신뢰성을 높인다. 세부 과정은 다음의 3단계로 구성된다.

(i) 정답 객체 Bbox 벡터 추출기(GT Object Bbox Vector Extractor)

첫 번째 단계는 전체 특징 맵에서 Ground-Truth(GT) 객체 영역에 해당하는 특징 벡터만을 공간적(spatial)으로 분리해내는 과정이다. 입력으로 주어지는 다중 스케일의 특징 맵을 F_{local}, F_{med}, F_{global}이라 하고, 각 스케일의 특징 맵 벡터를 $F_s \in R^{(H_S \times W_S \times C)}$라고 정의한다. 여기서 H_S와 W_S는 각 스케일의 높이와 너비, C는 트랜스포머 모델의 차원(본 논문 기준 256)이다.

먼저, GT Bbox가 원본 이미지 좌표계에 정의된 GT Bbox B_{gt} = {x, y, w, h}를 각 스케일의 공간 해상도 H_S, W_S에 맞춰 투영한다. 투영된 좌표를 기반으로 Bounding Box 내부 영역은 1, 외부 영역은 0의 값을 갖는 이진 공간 마스크(Binary Spatial Mask) M $\in \{0,1\}^{(H_S \times W_S \times 1)}$ 를 생성한다.

$$M_{box}^{i,j} = \begin{cases} 1, if\ pixel\ (i,j) \in B_{gt} \\ 0, otherwise \end{cases}$$

이 마스크 M_{box}는 특징 맵 F_s 와 요소별 곱(Element-wise Multiplication)을 수행하기 위한 필터로 사용되며, 이를 통해 Bounding Box 외부의 정보는 배제하고 내부의 특징 벡터들만을 추출한다.

(ii) 채널 별 L2 정규화(Channel-wise L2-norm)

추출된 Bounding Box 내부 영역이라 할지라도, 위장 객체의 특성상 객체(Foreground)와 배경(Background)이 혼재되어 있다. 따라서 픽셀별 중요도를 정량화하기 위해 '공간적 강도(Spatial Intensity)'를 측정한다. 본 논문에서는 특징 벡터의 크기(Magnitude)가 해당 위치의 정보량을 대변한다는 점에 착안하여, 채널 방향의 L2-norm을 활성도 스코어로 정의한다. 각 픽셀 위치 (i, j)에서의 특징 벡터 $v_{i,j} \in R^C$ 에 대하여, 활성도 스코어 $S_{i,j}$ 는 다음과 같이 계산된다.

$$S_{i,j} = \left\| v_{i,j} \right\|_2 = \sqrt{\sum_{c=1}^{C} \left(F_s^{(i,j,c)} \right)^2} \cdot M_{box}^{(i,j)}$$

이 과정을 통해 C차원의 특징 벡터는 하나의 스칼라 값 $S_{i,j}$로 압축되며, 전체 스코어 맵 S $\in R^{(H_S \times W_S \times 1)}$가 생성된다. 높은 $S_{i,j}$값은 해당 픽셀이 객체의 주요 특징(텍스처, 경계 등)을 강하게 표현하고 있음을 의미하며, 반대로 0에 가까운 값은 배경이거나 정보량이 적은 영역임을 나타낸다.

(iii) Top-K 요소 추출기(Top-K Component Extractor)

마지막 단계는 계산된 스코어 S를 기반으로 실제 객체를 대변하는 상위 K개의 핵심 성분(Component)을 최종 선별하는 과정이다. 여기서 선택되는 픽셀의 개수 K는 고정된 값이 아니며, 객체의 크기에 따라 유동적으로 결정되는 적응형 할당(Adaptive Allocation) 방식을 따른다.

GT 영역 내의 유효 픽셀 수(즉, M_{box}의 합)를 N_{fg}라고 할 때, 선택 비율 ρ는 객체 영역의 크기에 비례하여 결정된다. 본 연구에서는 다음과 같은 지수 함수적 증가 공식을 사용하여 작은 객체에 대해서는 보수적으로, 큰 객체에 대해서는 충분한 비율을 선택하도록 설계했다.

$$\rho = \rho_{min} + (\rho_{max} - \rho_{min}) \cdot (1 - e^{(-\frac{N_{fg}}{T})})$$

여기서 ρ_{min}, ρ_{max}는 각각 최소 및 최대 선택 비율이며, T는 감쇠 계수이다. 최종적으로 선택될 픽셀 수 K는 K = floor($N_{fg} \cdot \rho$)로 결정된다. 이후, 스코어 맵 S에서 상위 K개의 값을 가지는 픽셀의 인덱스를 추출하여 최종 Top-K 마스크 M_{topk}를 생성한다. 본 논문에서는 ρ_{min} = 0.1, ρ_{max} =0.18, T =96으로 설정하였다.

$$\Omega_{topk} = topk(S, K)$$

$$M_{topk}^{(i,j)} = \begin{cases} 1, if\ (i,j) \in \Omega_{topk} \\ 0, otherwise \end{cases}$$

이러한 Top-K 필터링 과정은 Bbox 내에 포함된 배경 노이즈를 효과적으로 제거하고, 가장 판별력(Discriminative)이 높은 객체 고유의 특징 영역만을 다음 단계인 프로토타입 정렬의 입력으로 전달하는 역할을 수행한다.

3.2.2 프로토타입 정렬 모듈 설계

3.2절의 관심 영역 추출기를 통해 선별된 객체의 핵심 픽셀들은 여전히 공간상에 흩어져 있는 상태이다. 프로토타입 정렬 모듈은 이 픽셀들의 정보를 집약하여 각 스케일(Local, Medium, Global)을 대표하는 하나의 '프로토타입 분포(Prototype Distribution)'로 변환하고, 스케일 간의 분포 차이를 최소화하는 손실(Loss)을 계산하는 역할을 수행한다. 단순한 평균값(Average Pooling)은 위장 객체의 복잡한 질감이나 패턴 정보를 충분히 반영하지 못하기 때문에, 본 모듈은 고차원 통계 정보를 활용하는 2차 프로토타입(Second-order Prototype)방식을 채택하여 정밀한 분포를 생성한다.

(i) 2차 프로토타입(Second-order Prototype)

Top-K 마스크 M_{topk}에 의해 선택된 픽셀들의 특징 벡터 집합을 $X \in R^{(K \times C)}$라고 하자. 여기서 K는 공간(spatial)에서 선택된 픽셀 수, C는 채널 수이다. 일반적인 Average Pooling은 X의 단순 평균만을 취하므로 특징 간의 상관관계를 무시한다. 이에 반해, 본 논문에서 제안하는 2차 프로토타입 방식은 특징 간의 공분산(Covariance) 정보를 반영하여 위장 객체의 미세한 식별 정보를 포착한다.

하지만 고차원 특징에 대한 공분산 행렬을 직접 계산하는 것은 연산량이 과도하여 본 논문에서는 Power Iteration 기법을 활용하여 효율적으로 주요 특징 성분을 추출한다. 초기 프로토타입 P_0는 채널별 평균으로 시작하며, 이후 반복적인 행렬 연산을 통해 정제된다.

$$P_0^{(c)} = Normalize\left(\frac{1}{K}\sum_{k=1}^{K} X_k^{(c)}\right), \qquad for\ c = 1, \ldots, C$$

그 후, T번의 반복(Iteration)을 통해 특징 벡터들 간의 상호작용을 반영하여 프로토타입을 업데이트한다. 각 단계 t에서의 업데이트 과정은 다음과 같다.

$$S_t = X \cdot P_t^T \ (Similarity\ Score)$$

$$P_{t+1}^{(c)} = Normalize(Softmax(S_t)^T \cdot X), \qquad for\ c = 1, \ldots, C$$

이 과정은 특징 벡터 X와 현재 프로토타입 P_t간의 유사도를 계산하고, 이를 가중치로 사용하여 X를 다시 결합하는 방식이다. 이를 통해 잡음은 억제되고 객체의 주된 특징 방향이 강화된 최종 2차 프로토타입 P_{final}을 얻게 된다. 본 논문에서는 특징맵 간 1차 프로토타입과 2차 프로토타입을 의미론적으로 정렬함으로써 특징맵을 효과적으로 융합한다.

(ii) 특징 분포 변환(Feature Distribution Transformation)

추출된 각 스케일 s ∈ {local, med, global}의 1, 2차 프로토타입 벡터는 C차원의 연속적인 값을 가진다. 이를 확률 분포 간의 거리 척도인 KL Divergence에 적용하기 위해서는, 각 채널의 값이 전체 특징에서 차지하는 상대적 중요도(확률)를 나타내도록 변환해야 한다. 따라서 Softmax 함수를 적용하여 프로토타입을 '특징 분포(Feature Distribution)' D_s로 변환한다.

$$D_s^{(c)} = \frac{e^{V_s^{(c)}/\tau}}{\sum_{j=1}^{C} e^{V_s^{(j)}/\tau}}$$

여기서 τ는 분포의 평탄도를 조절하는 온도(Temperature) 상수이다. 이 과정을 통해 D_s는 모든 채널의 합이 1이 되는 확률 분포의 성질을 갖게 되며, "어떤 채널(특징)이 해당 객체를 표현하는 데 더 중요한가"에 대한 정보를 담게 된다.

(iii) 전역 중심 분포 정렬 손실(Global-centric Distribution Alignment Loss)

마지막으로, 3.2절을 통해 추출된 각 스케일의 특징 분포가 일관된 의미론적 표현을 갖도록 강제하는 손실 함수를 설계한다. 기존의 인접 스케일 간 정렬 방식은 오차가 누적될 수 있다는 단점이 있다. 따라서 본 논문에서는 가장 넓은 Receptive Field을 가지며 객체의 문맥 정보를 가장 완벽하게 포함하고 있는 Global 스케일(D_{global})을 정답지(Anchor)로 설정한다.

이에 따라 하위 스케일인 Local(D_{local})과 Medium(D_{med}) 분포가 모두 Global 분포를 직접 따르도록 강제하는 전역 중심 정렬(Global-centric Alignment)방식을 채택한다. 손실 함수로는 분포 간의 차이를 측정하는 데 가장 우수한 성능을 보인 KL Divergence를 사용한다.

첫째, 고해상도의 세밀한 정보를 담고 있는 Local 스케일(D_{local})을 Global 스케일(D_{global})에 직접 정렬시킨다. 이는 로컬 특징이 지엽적인 정보에 매몰되지 않고, 전역적인 문맥을 반영하도록 유도한다.

$$L_{local \to global} = \mathcal{D}_{KL}(D_{global} \,||\, D_{local}) = \sum_{c=1}^{C} D_{global}^{(c)} \, log \frac{D_{global}^{(c)}}{D_{local}^{(c)}}$$

둘째, 중간 단계의 정보를 담고 있는 Medium 스케일(D_{med}) 또한 Global 스케일(D_{global})에 정렬시킨다. 이는 중간 특징 맵이 상위 의미론적 정보와 일치하도록 보정한다.

$$L_{medium \to global} = \mathcal{D}_{KL}(D_{global} \,||\, D_{medium}) = \sum_{c=1}^{C} D_{global}^{(c)} \, log \frac{D_{global}^{(c)}}{D_{medium}^{(c)}}$$

최종적인 Distribution Loss (L_{dist})는 이 두 손실의 합으로 정의된다. 이를 통해 모델은 모든 스케일에서 Global 스케일 기준의 통일된 특징 분포를 학습하게 된다.

$$L_{dist} = L_{local \rightarrow global} + L_{med \rightarrow global}$$

이 손실 함수는 역전파를 통해 Local, Medium 특징맵의 1x1 합성곱층을 업데이트한다. 결과적으로, 모델은 스케일이나 해상도의 변화와 무관하게 Global Anchor가 제공하는 강력한 의미론적 가이드라인을 따르게 되어, 위장 객체 탐지 시 발생하는 다중 스케일 간의 의미론적 불일치(Semantic Inconsistency) 문제를 근본적으로 해결한다.

4. 실험 및 분석

본 장에서는 3 장에서 제안한 방법에 대한 실험을 진행한다. 4.1 절에서는 실험 환경과 실험에서 사용한 데이터셋에 대해서 설명하고, 4.2 절에서는 실험 및 성능 분석 결과, 4.3 절에서는 관련 연구와의 비교를 진행한다. 4.4 절에서는 제안한 모델의 한계에 대해 분석한다.

4.1 실험 환경 및 데이터셋

본 논문에서 진행된 실험은 GPU Nvidia RTX 3090 24GB 2장과 Ubuntu 20.04 환경에서 Pytorch 프레임워크를 이용하여 구현하였다.

본 연구에서는 제안하는 Top-K 프로토타입 정렬 기법의 위장 객체 탐지 성능을 검증하고, 일반 객체 탐지 모델에 미치는 영향을 종합적으로 분석하기 위해 다양한 데이터셋을 활용하여 실험을 진행하였다. 실험에 사용된 데이터셋은 크게 위장 객체 탐지 데이터셋과 일반 객체 탐지 데이터셋으로 분류된다.

첫째, 위장 객체 탐지 성능 검증을 위한 주 데이터셋으로는 MHCD2022[10]와 COD10K[4]를 사용하였다. MHCD2022는 군사적 위장 환경에 특화된 데이터셋으로, 탱크, 위장복을 입은 보병 등 본 연구의 핵심 목표인 국방 감시 시스템 적용에 가장 적합한 객체들로 구성되어 있다. 반면 COD10K는 자연환경 속의 은폐된 동물 등을 포함하는 가장 대표적인 위장 객체 탐지 벤치마크이다. 다만, COD10K는 본래 픽셀 단위의 마스크 정보를 제공하는 Segmentation 데이터셋이므로, 본 연구의 탐지 모델 학습을 위해 이를 Bounding Box 형태로 변환하는 전처리 과정을 수행하였다. 이때 박스의 좌표는 객체 마스크의 최외각 지점을 기준으로 생성하였으며, 곤충의 더듬이나 다리 끝과 같이 미세하게 돌출된 부분까지 모두 박스 경계 내부에 포함되도록 정의하였다.

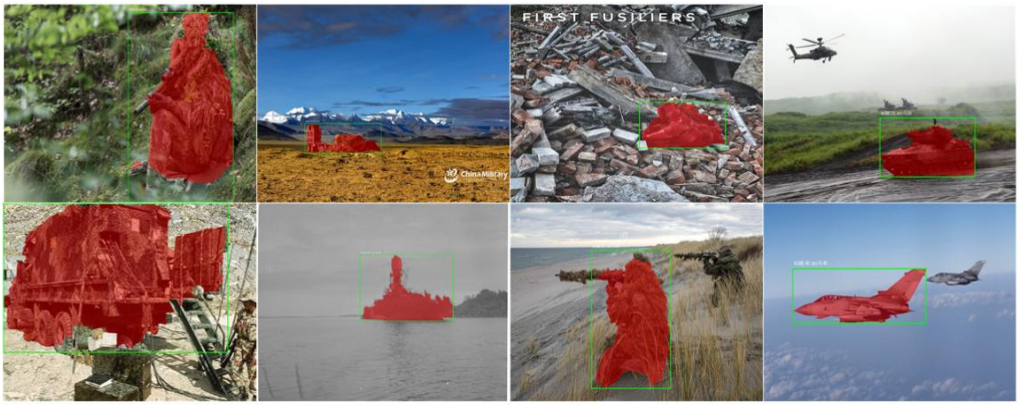

그림 2 MHCD2022 데이터셋

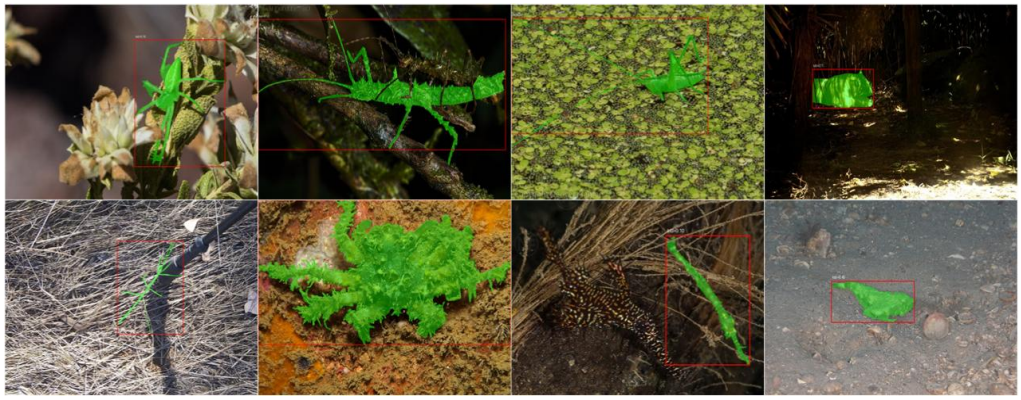

그림 3 COD10K 데이터셋

특히, 본 연구에서는 COD10K 데이터셋의 특성을 고려하여 별도의 정제 과정을 거친 부분 데이터셋을 추가로 구성하였다. 2.3절에서 논의한 바와 같이, 기존 COD10K 데이터셋은 Bounding Box 대비 실제 객체가 차지하는 비율(Bbox-to-Object Ratio)이 매우 낮은 데이터가 다수 포함되어 있다. 이러한 데이터는 Bbox 내에 과도한 배경 정보(Background Noise)를 포함하여 특징 정렬 학습을 방해하는 요인이 된다. 따라서 본 실험에서는 MHCD2022와의 비교 분석 및 Bbox 기반 탐지 모델의 적합성을 검증하기 위해, Bbox와 객체 비율이 평균 0.5 이상이 되도록 COD10K 데이터 중 1,616장을 선별하여 별도의 실험셋(아래 표 COD10K_0.5)

을 구성하였다. Bbox와 객체 비율을 계산하기 위해 Segmentation 모델인 Segment Anything[23] 모델을 사용해서 각 데이터셋의 Bbox 내 Segmentation 마스크를 구하고, Bbox-Obj 비율을 계산해서 평균을 산출하였다.

둘째, 제안하는 방법이 일반적인 객체 탐지 성능에 미치는 범용적 효과성을 검증하기 위해 Pascal VOC 2007[24]과 MSCOCO 2017[3] 데이터셋을 활용하였다. 실험의 효율성과 사용 데이터셋의 절대량을 맞추기 위해 전체 데이터셋을 사용하는 대신, Pascal VOC 2007 데이터 5,000장과 MSCOCO 2017 Validation 데이터 5,000장을 사용하였다.

표 1 활용 데이터셋의 세부 특징

데이터셋	데이터 수	클래스 수	Bbox-Obj 비율 평균
MHCD2022[10]	3,000	5	0.51
COD10K[4]	5,066	2(객체, 배경)	0.38
COD10K_0.5	1,616	2(객체, 배경)	0.51
MSCOCO[3]	5,000	80	0.56
Pascal VOC[24]	5,011	20	0.52

모든 실험 데이터셋은 모델의 공정한 학습 및 평가를 위해 학습 데이터(Train set)와 검증 데이터(Validation)를 8:2의 비율로 무작위 분할(Random Split)하여 사용하였다. 이를 통해 제안 모델이 위장 객체뿐만 아니라 다양한 도메인의 객체에 대해 가지는 탐지 성능과 특징 정렬의 효과를 다각도로 분석하였다.

4.2 실험 및 성능 분석

4.2.1 위장 객체 탐지 성능 분석

본 절에서는 제안하는 방법이 위장 객체 탐지 모델의 성능에 미치는 영향을 정량적 수치와 정성적 시각화 자료를 통해 심도 있게 분석한다.

(i) 데이터셋 특성에 따른 성능 및 원인 분석

먼저, 본 연구의 핵심 적용 대상인 군사적 위장 환경을 대변하는 MHCD2022[10] 데이터셋에 대한 실험 결과를 분석하였다. <표 2>에 제시된 바와 같이, 기준 모델인 D-FINE[9] (Baseline)과 비교하여 제안 모델은 모든 주요 지표에서 뚜렷한 성능 향상을 달성하였다. 구체적으로, 객체의 존재 여부를 판단하는 mAP@0.5 지표는 75.27%에서 77.94%로 2.67%p 상승하였으며, 더욱 정밀한 위치 탐지를 요구하는 mAP@0.75 지표는 55.96%에서 61.49%로 5.53%p의 큰 폭으로 상승하였다. 또한, 다양한 IoU 임계값을 종합적으로 고려한 mAP@[0.5:0.95] 지표 역시 51.71%에서 54.95%로 3.24%p 향상되었다. 이는 제안하는 전역 특징 정렬 기법이 배경과 매우 유사하여 식별이 어려운 군사 객체의 특징 표현을 효과적으로 강화하고, 위장된 객체의 경계를 더욱 명확하게 구분하는 데 기여했음을 시사한다.

표 2 MHCD2022 데이터셋을 활용한 제안 모델의 성능

모델 (data:MHCD2022)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
Faster-RCNN[25]	53.56	33.53	32.03
DETR[15]	56.58	36.22	34.71
MHNet[10]	56.76	38.16	36.89
D-FINE S[9] (baseline)	75.27	55.96	51.71
Ours	77.94(+2.67)	61.49(+5.53)	54.95(+3.24)

반면, <표 3>에서와 같이 자연 환경의 위장 객체를 포함하는 COD10K[4] 데이터셋 전체를 활용한 초기 실험에서는 MHCD2022와 상반된 결과가 나타났다. 제안 기법을 적용했음에도 불구하고 성능 향상이 미미하거나, 일부 지표에서는 오히려 베이스라인 모델과 유사한 수준에 머무르는 현상이 관찰되었다. 이러한 성능 불일치의 원인을 규명하기 위해, 학습 과정에서 모델이 주목하는 영역을 시각화한 히트 맵(Heatmap) 분석을 수행하였다.

표 3 COD10K 데이터셋을 활용한 제안 모델의 성능

모델 (data:COD10K)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
D-FINE S[9] (baseline)	76.81	48.92	47.73
Ours	77.81(+1.0)	47.72(-1.2)	47.63(-0.1)

<그림 4, 5>은 MHCD2022와 COD10K 데이터셋에 대해 충분한 학습 스텝(Step)이 진행된 후, Local 특징 맵의 활성화(Activation) 양상을 비교한 것이다. MHCD2022의 경우 학습이 진행됨에 따라 Local 특징 맵이 객체 영역에 강하게 집중되며 배경과의 대비가 뚜렷해지는 경향을 보였다. 그러나 COD10K의 경우, 충분한 학습이 이루어졌음에도 불구하고 활성화 영역이 객체와 배경을 명확히 구분하지 못하거나, 심지어 객체 영역의 활성도가 초기보다 감소하는 역설적인 현상이 확인되었다.

학습 전

학습 후

(5000 step)

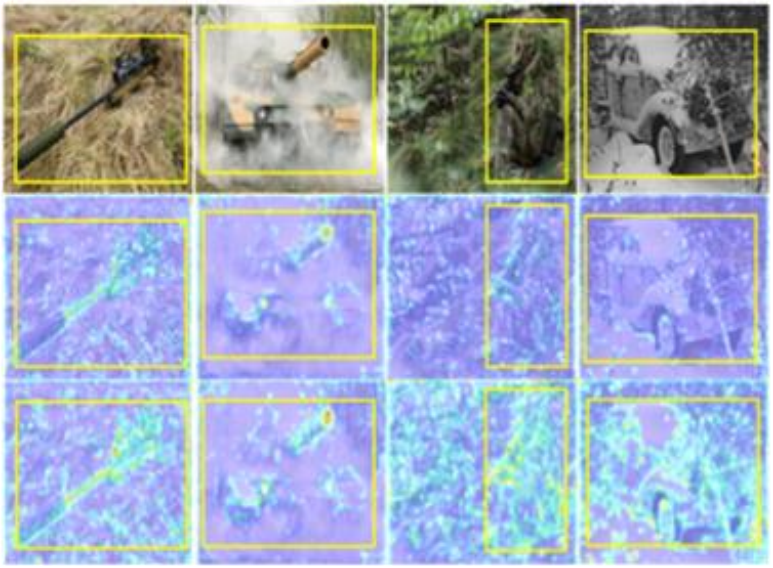

그림 4 제안한 모델의 Local 특징 맵의 히트 맵(MHCD2022)

학습 전

학습 후

(5000 step)

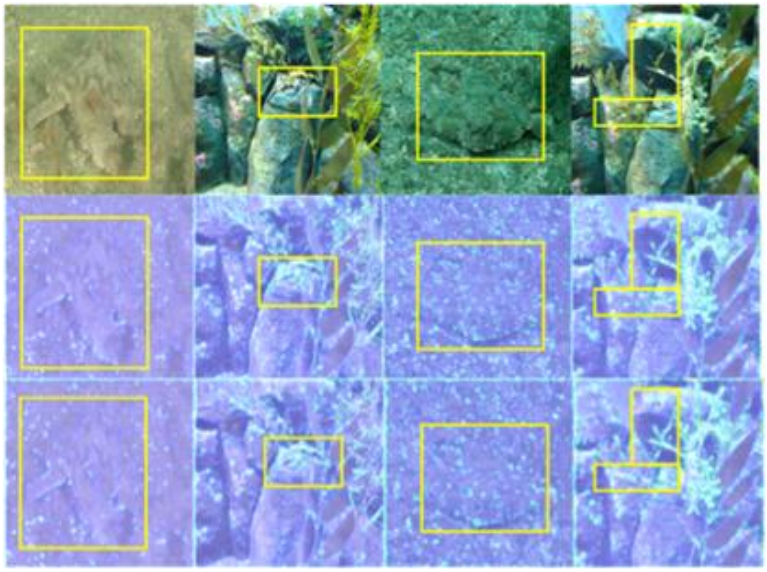

그림 5 제안한 모델의 Local 특징 맵의 히트 맵(COD10K)

본 연구에서는 이러한 현상의 근본적인 원인이 Bbox 내 객체가 차지하는 비율(Bbox-Obj Ratio)의 구조적 차이에 있다고 가정하였다. 두 데이터셋의 객체 마스크 비율을 정밀 분석한 결과, MHCD2022의 데이터는 전차나 차량 등 부피가 큰 객체 위주로 구성되어 대부분 0.4 이상의 높은 객체 점유율을 보였다. 반면, COD10K는 곤충의 더듬이나 얇은 다리 등 가늘고 긴 객체가 다수 포함되어 있어, Bbox의 크기는 크지만 실제 객체가 차지하는 픽셀 영역은 0.1 미만인 데이터가 상당수 존재함을 확인하였다.

그림 6 Bbox 내 객체의 비율이 낮은 예시(COD10K)

객체 비율이 극도로 낮은 데이터의 경우, Bbox 내부 정보의 대부분이 객체가 아닌 배경으로 구성된다. 이러한 상황에서 별도의 필터링 없이 특징 정렬을 수행하게 되면, 모델은 객체의 고유 특징이 아닌 배경의 노이즈(Noise)를 학습하게 된다. 즉, 배경 정보가 전역 프로토타입에 강제로 정렬되면서, 모델이 배경을 객체로 오인하거나 객체 식별력을 상실하게 만드는 부정적인 효과를 초래한 것이다.

이 가설을 검증하기 위해 MSCOCO[3], Pascal VOC[24] 데이터셋의 평균 Bbox-Obj 비율(0.5 이상)과 유사하도록 COD10K 데이터셋 중 객체 비율이 높은 순서대로 1,616장을 선별하여 재실험을 수행하였다. 그 결과, 정제된 데이터셋에서는 앞서 관찰되지 않았던 뚜렷한 성능 향상이 확인되었으며, MHCD2022 실험 결과와 유사한 성능 개선 추이를 보였다<표 4, 그림 7>. 이는 Bbox 기반의 위장 객체 탐지 접근법에서 객체 내부 정보의 밀도(Density)가 학습 성능을 결정짓는 핵심 요소임을 입증하며, 배경 노이즈를 배제하고 순수한 객체 특징만을 추출하는 과정이 필수적임을 강력하게 시사한다.

표 4 COD10K_0.5 데이터셋을 활용한 제안 모델의 성능

모델 (data:COD10K_0.5)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
D-FINE S[9] (baseline)	91.51	54.52	54.63
Ours	93.21(+1.7)	63.62(+9.1)	60.03(+5.4)

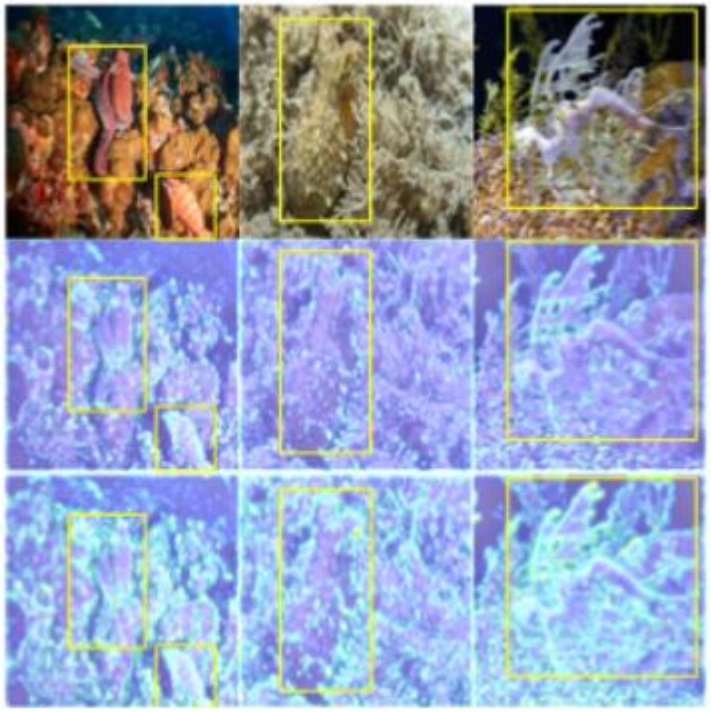

그림 7 데이터셋 조정 후 Local 특징 맵의 활성화 맵(COD10K_0.5)

(ii) 구성 요소별 유효성 검증(Ablation Study)

앞선 원인 분석을 바탕으로, 제안 모델의 핵심 구성 요소인 Top-K 전략, 앵커 선정 기준, 2차 프로토타입 방식, 그리고 손실 함수에 대한 상세한 유효성 검증을 수행하였다.

첫째, Top-K 위치 추출 전략의 효과를 검증하기 위해, Bbox 내 모든 픽셀을 정렬 대상으로 삼는 'All Bbox' 방식과 성능을 비교하였다. <표 5>의 실험 결과에 따르면, 'All Bbox' 방식은 Baseline보다는 성능이 좋았으나, mAP@0.5 기준 76.14%를 기록하여 제안 모델(77.9%)에 비해 현저히 낮은 성능을 보였다. 이는 배경 노이즈가 다수 포함된 전체 영역을 무분별하게 정렬하는 방식이 오히려 모델의 식별력을 저해할 수 있음을 의미한다. 반면, 유의미한 특징을 가진 상위 K개의 픽셀만을 동적으로 선별하여 정렬하는 본 연구의 Top-K 전략은 배경 잡음을 효과적으로 억제하고 객체 특징을 강화하여 성능을 극대화하는 데 필수적인 역할을 수행함을 확인하였다.

표 5 제안한 Top-K 방법과 Bbox 내 모든 요소를 강화하는 방법의 성능 차이

모델 (data:MHCD2022)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
D-FINE S[9] (baseline)	75.27	55.96	51.71
Top_k(Ours)	77.94	61.49	54.95
All_Bbox	76.14	60.62	54.90

둘째, 특징 정렬의 기준점이 되는 앵커(Anchor) 특징 맵 선정에 따른 성능 변화를 분석하였다. 실험 결과, 3장에서 설계한 바와 같이 트랜스포머 인코더(Transformer Encoder)를 통과하여 이미지 전체의 전역적 문맥(Global Context)을 가장 풍부하게 함유하고 있는 Global 특징 맵을 앵커로 설정했을 때 가장 우수한 성능을 기록하였다.<표 6> 반면, 수용 영역(Receptive Field)이 좁아 지엽적인 정보에 치우친 Local 특징 맵이나, 중간 단계인 Medium 특징 맵을 앵커로 사용할 경우, 전체적인 특징 분포를 이끄는 강력한 기준점이 부재하여 성능 하락이 발생하였다. 이는 하위 스케일의 특징들을 상위 스케일의 전역 정보에 일치시키는 전역 중심(Global-centric) 정렬 전략의 타당성을 입증한다.

표 6 Anchor 특징 맵에 따른 성능 차이

모델 (data:MHCD2022)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
Global anchor (Ours)	77.94	61.49	54.95
Medium anchor	76.70	59.67	53.59
Local anchor	76.16	58.40	53.96

셋째, 추출된 픽셀 정보를 하나의 벡터로 압축하는 프로토타입 집계 방식에서는 단순 평균(Channel-wise Average)만을 사용하는 방식보다 추가로 Power Iteration을 활용한 방식이 더 높은 성능을 기록하였다<표 7>. 단순 평균은 위장 객체의 복잡한 패턴을 일차원적으로 요약하여 정보 손실이 발생하는 반면, Power Iteration은 특징 간의 공분산(Covariance)과 상호작용을 반영함으로써 위장 객체 내부에서도 가장 주도적이고 차별화된 특징 방향을 효과적으로 포착해내기 때문으로 분석된다. 반면, Power Iteration을 사용하지 않고 채널 간 공분산 행렬(Covariance Matrix)을 직접 산출하여 고유값 분해(Eigen Decomposition)를 수행하는 방식은 치명적인 연산 비효율성을 보였다. 특징 채널 수를 C라고 할 때, 공분산 행렬을 구축하고 분해하는 과정은 $O(c^2)$ 이상의 시간 복잡도와 공간 복잡도를 요구한다. 이는 다중 스케일 특징 맵을 처리해야 하는 본 모델의 학습 과정에서 GPU 메모리 오버헤드(Memory Overhead)를 급격히 증가시켜, 학습이 불가능하였다. 이에 비해 제안하는 Power Iteration 방식은 O(C)의 선형 복잡도만으로도 주성분 벡터를 효과적으로 근사할 수 있어, 실시간 모델의 학습 효율성과 성능을 동시에 만족시키는 최적의 선택임을 확인하였다.

마지막으로, 특징 분포 간의 일치도를 측정하는 손실 함수(Loss Function)비교 실험을 진행하였다. <표 7>에 나타난 바와 같이, 단순한 값의 차이를 측정하는 L1 Loss(75.73%)나 방향성만을 고려하는 Cosine Similarity(77.77%)에 비해, 확률 분포 간의 정보량 차이를 측정하는 KL Divergence가 77.94%(mAP@0.5) 및 61.49%(mAP@0.75)로 가장 우수한 성능을 달성하였다. 이는 다중 스케일 특징 맵을 확률 분포로 변환하여 정렬할 때, KL Divergence가 분포의 형상과 불확실성을 가장 정밀하게 최적화하여 스케일 간 의미론적 일관성을 강화하기 때문이다.

표 7 평균과 주성분, loss 변경에 따른 성능 차이

모델(data:MHCD2022)		mAP@0.5	mAP@0.75	mAP@[0.5:95]
평균 + 주성분	Cosine similarity	77.77	60.22	54.88
	L1 loss	75.73	59.59	53.67
	KL divergence (Ours)	77.94	61.49	54.95
평균	Cosine similarity	76.52	57.30	53.34
	L1 loss	75.00	61.23	54.22
	KL divergence	77.42	61.72	55.17

결론적으로, 본 실험들은 객체 비율을 고려한 Top-K 기반의 특징 추출, Global Anchor 중심의 정렬 구조, 평균과 Power iteration을 통한 2차 프로토타입 생성, 그리고 KL Divergence 손실 함수의 유기적인 결합이 실시간 위장 객체 탐지 성능을 극대화하는 최적의 구성임을 입증하였다.

4.2.2 일반 객체에 대한 효과성 분석

본 절에서는 제안하는 Top-K 프로토타입 정렬 기법이 객체의 경계가 뚜렷한 일반 객체 탐지(General Object Detection) 환경에서는 어떠한 영향을 미치는지 분석하기 위해, Pascal VOC 2007[24]과 MSCOCO 2017[3] 데이터셋을 활용하여 검증을 수행하였다.

<표 8, 9>는 두 데이터셋에 대한 베이스라인 모델(D-FINE[9])과 제안 모델의 성능 비교 결과를 나타낸다. 실험 결과, 위장 객체 데이터셋(MHCD2022[10])에서 유의미한 성능 향상을 보였던 것과 대조적으로, 일반 객체 데이터셋에서는 베이스라인 대비 소폭의 성능 저하가 관찰되었다. 구체적으로 Pascal VOC 데이터셋에서 제안 모델의 mAP@0.5는 74.80%를 기록하여 베이스라인(75.47%) 대비 약 0.67%p 하락하였으며, mAP@[0.5:0.95] 지표 역시 54.21%로 베이스라인(54.27%)과 유사하거나 미세하게 낮은 수치를 보였다. 이러한 경향은 MSCOCO 데이터셋에서도 동일하게 나타나, mAP@0.5 기준 41.96%에서 40.73%로, mAP@[0.5:0.95] 기준 27.93%에서 26.98%로 감소하였다.

표 8 제안한 방법을 활용한 일반 객체 데이터셋에 대한 성능(Pascal VOC)

모델 (data:Pascal VOC)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
D-FINE S[9] (baseline)	75.47	58.37	54.27
Ours	74.80	58.72	54.21

표 9 제안한 방법을 활용한 일반 객체 데이터셋에 대한 성능(MSCOCO)

모델(data:MSCOCO)	mAP@0.5	mAP@0.75	mAP@[0.5:95]
D-FINE S[9] (baseline)	41.96	29.28	27.93
Ours	40.73	28.27	26.98

이러한 결과는 일반 객체와 위장 객체의 시각적 특성 차이와 제안 기법의 설계 목적에서 기인한 것으로 분석된다. 위장 객체(Camouflaged Object)는 배경과 전경의 텍스처가 매우 유사하여 특징 맵 상에서 경계가 불분명하고(Feature Ambiguity), 스케일 간의 정보 불일치가 빈번하게 발생한다. 따라서 본 연구가 제안한 '전역 특징 정렬'은 이러한 모호한 특징 분포를 강제로 일치시켜 식별력을 강화하는 데 결정적인 역할을 수행한다.

반면, Pascal VOC[24]나 MSCOCO[3]에 포함된 일반 객체들은 배경과의 색상 및 명암 대비(Contrast)가 뚜렷하고 경계가 명확하다. 이러한 데이터에서는 기존의 CNN이나 트랜스포머 기반 백본 네트워크가 이미 충분히 강건한 특징을 추출하고 융합하고 있다. 즉, 이미 객체와 배경이 특징 공간상에서 잘 분리되어 있는 상태이므로, 제안하는 정렬 기법을 통해 분포를 추가적으로 제약하는 것은 불필요한 연산이 되거나, 오히려 과도한 정규화로 작용하여 최적화된 특징 표현을 저해할 수 있음을 시사한다.

결론적으로, 본 실험 결과는 제안하는 Top-K 프로토타입 정렬 기법이 모든 객체 탐지 문제에 범용적으로 성능을 높이는 만능 해결책이 아니라, 배경과 객체의 구분이 모호하여 특징 추출이 어려운 '위장 객체 탐지'와 같은 고난이도(Hard) 과제에 특화된 솔루션임을 입증한다. 따라서 명확한 경계를 가진 일반 객체보다는, 국방 감시나 은폐된 표적 식별과 같은 특수 목적 환경에서 그 효용성이 극대화됨을 확인할 수 있다.

4.3 관련 연구와의 비교

본 연구에서 제안한 'Top-K 프로토타입 정렬 기반의 실시간 위장 객체 탐지 모델'은 기존의 위장 객체 탐지(COD) 연구 및 최신 객체 탐지 기법들과 비교하여 구조적 효율성과 특징 처리 방식에서 분명한 차별성을 갖는다. 본 절에서는 Segmentation 기반 모델, 일반 객체 탐지 모델, 그리고 프로토타입 기반 연구들과의 비교를 통해 제안 방법의 학술적 기여를 구체적으로 논한다.

1) 최신 Segmentation 기반 COD 모델 및 MHCD2022[10] 벤치마크와의 비교

현재 위장 객체 탐지(COD) 분야의 최신 연구인 SINet[4], ZoomNet[19] 등은 객체의 정밀한 형상 복원을 위해 다단계로 특징을 정제하거나 점진적으로 스케일을 확장하는 방식을 채택하고 있다. 이러한 Segmentation 기반의 접근은 픽셀 단위의 정밀한 마스크(Mask)를 생성하여 객체의 윤곽을 완벽하게 복원한다는 장점이 있다. 하지만 입력 이미지의 해상도 전체에 대해 조밀한(Dense) 예측을 수행해야 하므로, 모델의 구조가 깊어지고 연산량이 급증하는 근본적인 한계를 가진다. 이는 0.1초의 지연도 치명적인 국방 경계 시스템이나 고속 드론 감시 체계에 적용하기에는 높은 진입 장벽으로 작용한다. 또한, 본 연구에서 활용한 MHCD2022 데이터셋과 함께 제안된 벤치마크 모델인 MHNet(Military High-level Camouflage object detection Network)[10]은 인간의 시각 인지 과정을 모방하기 위해 고안되었다. MHNet은 전역적 특징을 탐색하는 SPG(Subject Perception Gathering) 모듈과 부분-객체 간의 관계를 파악하는 PRM(Part-object Relationships Mining) 모듈 등 다수의 복잡한 서브 모듈을 결합하여 은폐된 객체를 픽셀 단위로 식별한다. 이러한 다단계 구조는 정밀한 형상 복원에는 유리할 수 있으나, Bbox 기반의 모델임에도 과도한 연산량을 유발하여 국방 경계 시스템이 필수적으로 갖춰야 할 실시간성을 확보하는 데에는 구조적인 한계가 존재한다.

표 10 객체 탐지 모델 별 추론 시간 및 지연율

모델	FPS	지연율(ms)
Faster-RCNN[25]	15~23	45~65
DETR[15]	10~20	50~80
MHNet[10]	10~15	65~100
D-FINE S[9] (Ours)	75~130	8~13

이에 반해, 본 연구는 MHCD2022 데이터셋의 객체(전차, 인원 등)가 비교적 정형화된 형태를 가지며 Bbox 내 객체 점유율이 높다는 점에 주목하여, 실시간 처리에 최적화된 Bbox 기반 탐지 방식을 채택하였다. D-FINE[9] 모델을 기반으로 한 본 제안 방법은 실시간 객체 탐지 모델의 구조는 유지하되, 위장 객체에 효과적인 학습 방법을 활용해 성능 면이나, 속도 면에서 높은 성능을 달성하였다. 또한, 제안하는 방법은 학습 시에도 많은 추가 자원을 요구하지 않아 효율적이다. 베이스라인 모델과 제안 모델의 MHCD2022 데이터셋 학습시간을 10 epoch 동안 추적한 결과, 학습에 소요되는 시간은 평균 34.2초에서 40.2초로 증가하였으며, 이는 적은 컴퓨팅 자원만을 가지고, 객체 탐지 모델을 위장객체에 효과적인 방향으로 학습할 수 있음을 의미한다.

결과적으로 기존 벤치마크 모델들이 '형상 복원'이나 '높은 성능'을 위해 연산 효율성을 희생했다면, 본 연구는 학습 단계에 단순 loss 추가만으로 추론 시 추가 연산 없이 실시간 탐지 속도와 높은 성능이라는 두 가지 실질적인 목표를 동시에 달성하였다.

2) 일반 객체 탐지 및 정렬 기법과의 비교

Align-DETR[21]이나 본 연구의 베이스라인인 D-FINE과 같은 최신 객체 탐지 모델들은 특징 맵의 정렬(Alignment)을 통해 분류와 회귀 간의 불일치를 해결하는 데 집중한다 . 그러나 이러한 접근은 객체와 배경의 대비가 명확하여, 국소적인 특징만으로도 객체 식별이 가능하다는 가정을 전제로 한다. 따라서 배경과 특징이 혼재된 위장 객체 환경에서는 지역적 특징 강화만으로는 한계가 발생한다.

본 연구의 차별점은 하위 스케일의 불확실한 특징을 상위 스케일의 확실한 전역 문맥(Global Context)에 강제로 일치시키는 '전역 중심(Global-centric) 정렬' 구조에 있다. 일반 탐지 모델이 특징의 위치를 정제하는 데 그친다면, 제안 모델은 특징의 분포 자체를 전역 정보에 맞게 재구성함으로써 모호한 위장 패턴을 문맥적으로 해석할 수 있도록 유도한다 . 이는 단순한 정렬을 넘어, 위장 객체라는 특수한 도메인에 특화된 학습 전략이라 할 수 있다.

3) 프로토타입(Prototype) 활용 방법들과의 비교

기존 컴퓨터 비전 분야에서 프로토타입을 활용하는 연구들은 크게 클래스 단위의 전역 프로토타입을 활용하는 방식과 확률적 할당을 활용하는 방식으로 나눌 수 있으며, 본 연구는 이들과 구조적 효율성 및 노이즈 강건성 측면에서 차별화된다.

첫째, 구조적 효율성 측면이다. 일반적인 프로토타입 기반의 Segmentation이나 Few-shot Learning 연구들은 데이터셋 전체의 클래스 특징을 대표하는 프로토타입 벡터를 학습하기 위해 별도의 메모리 뱅크(Memory Bank)를 구축하거나, K-means와 같은 클러스터링 알고리즘을 수행하여 프로토타입을 지속적으로 업데이트한다. 이러한 방식은 프로토타입 저장을 위한 추가적인 파라미터 공간을 요구하며, 학습 과정에서 클러스터 중심을 맞추기 위한 연산 비용이 발생한다.

반면, 본 연구의 제안 기법은 전역적인 클래스 정보를 저장하는 대신, 각 이미지의 Bbox 내부에서 즉각적으로 생성되는 '인스턴스 단위 프로토타입(Instance-level Prototype)'을 활용한다. 이는 Bbox 내의 Top-K 특징만을 추출하여 동적으로 프로토타입을 구성하므로, 별도의 저장 공간이나 학습 가능한 파라미터(Learnable Parameters)가 전혀 필요하지 않다는 구조적 이점을 가진다.

둘째, 노이즈에 대한 강건성(Robustness) 측면이다. 최근 약지도 학습(Weakly Supervised Learning) 분야의 POT[22]나 PANet[7] 등의 연구는 특징 맵의 전체적인 분포를 고려하여 최적 수송(Optimal Transport)과 같은 확률적 방식으로 프로토타입을 할당한다. 이러한 확률적 접근은 정보량이 풍부한 일반 이미지에서는 효과적이나, Bbox 내 배경 비율이 극도로 높은 위장 객체 데이터에서는 배경 노이즈까지 가중치에 반영하여 오히려 식별력을 떨어뜨리는 역효과를 낳는다.

이와 대조적으로 본 연구의 Top-K 전략은 확률적 가중치가 아닌 물리적인 선별 방식을 취한다. 이는 배경일 확률이 높은 하위 활성 픽셀들을 원천적으로 배제하고, 가장 신뢰도 높은 상위 K개의 정보만을 사용하여 프로토타입을 구축한다. 즉, '모든 정보를 고려하는 것'보다 '확실한 정보만 남기는 것'이 노이즈가 압도적으로 많은 위장 객체 탐지 환경에서 더욱 강건한 특징 표현을 가능케 함을 입증하였다.

표 11 프로토타입 방법 별 특징 및 한계/차별점

구 분	방법론 및 특징	한계점 및 차별점
전역 프로토타입 방법 (PANet[7])	클래스의 특징을 대표하는 프로토타입 벡터 학습	프로토타입 저장을 위한 추가적인 파라미터 공간 필요
	메모리 뱅크 또는 클러스터링으로 지속 업데이트	클러스터 중심을 맞추기 위한 추가 연산 비용 발생
확률적 할당 방법 (POT[22])	특징 맵의 전체적 분포를 고려해 할당	배경비율이 높은 위장 객체에 대해서는 배경 노이즈까지 가중치에 반영 위험
	최적 수송과 같은 확률적 가중치 방식 적용	위장 객체에 적용 시 객체 식별력이 떨어질 수 있음
제안 방법	Bbox 내 상위 K개의 신뢰도 높은 특징 추출	별도 저장 공간이나 학습 파라미터가 필요 없음
	이미지마다 즉각적으로 추출된 인스턴스 프로토타입만을 사용	배경일 수 있는 픽셀을 물리적으로 배제함으로써 노이즈에 강건함

5. 결론 및 향후 과제

본 논문에서는 국방 및 감시 시스템과 같이 즉각적인 대응이 요구되는 환경에서 위장된 객체를 정밀하게 탐지하기 위한 '객체 마스크 단위 전역 특징 정렬' 기법을 제안하였다. 기존의 위장 객체 탐지(COD) 연구가 연산량이 높은 Segmentation 방식에 치중되어 실시간성 확보에 어려움이 있었다면, 본 연구는 실시간 탐지가 가능한 Bounding Box 기반의 D-FINE[9] 모델을 채택하고, 이에 최적화된 새로운 학습 전략을 도입하여 속도와 정확도라는 두 마리 토끼를 동시에 잡고자 하였다.

본 연구의 핵심 기여점은 다음과 같이 요약할 수 있다.

첫째, 위장 객체의 모호한 경계와 Bounding Box 내부의 배경 노이즈 문제를 해결하기 위해 관심 영역 추출기(Region of Interest Extractor)를 설계하였다. 이를 통해 객체 영역 내에서도 가장 유의미한 상위 K개의 픽셀만을 동적으로 선별하여 특징 정렬의 대상으로 삼음으로써, 배경 정보의 혼입을 최소화하고 객체 고유의 특징 표현을 강화하였다.

둘째, 다중 스케일 특징 맵 간의 의미론적 불일치 문제를 해결하기 위해 전역 중심 분포 정렬(Global-centric Distribution Alignment) 학습 방법을 제안하였다. 트랜스포머 인코더를 통과하여 전역적 문맥 정보를 함유한 Global 특징 맵을 앵커(Anchor)로 설정하고, Power Iteration을 통해 추출된 2차 프로토타입 분포를 KL Divergence 손실 함수로 정렬시켰다. 이를 통해 하위 스케일의 특징들이 상위 스케일의 일관된 문맥을 따르도록 유도하여 탐지 성능을 극대화하였다.

셋째, 제안 모델의 효용성을 검증하기 위해 군사적 위장 객체 데이터셋인 MHCD2022[10]와 일반 위장 데이터셋인 COD10K[4]에 대해 광범위한 실험을 수행하였다. 실험 결과, 군사적 객체(탱크, 위장 보병 등)에 대해 mAP@0.5 기준

2.67%p, mAP@0.75 기준 5.53%p의 유의미한 성능 향상을 달성하였으며, 이는 본 연구가 목표로 한 국방 감시 시스템으로서의 실전적 가치를 입증한다. 또한, COD10K 데이터셋 분석을 통해 Bounding Box 기반 탐지에서 'Bbox 대비 객체 비율'이 학습 성능에 미치는 결정적인 영향을 규명하였다.

하지만 본 연구는 다음과 같은 몇 가지 한계점을 가지며, 이는 향후 연구를 통해 개선되어야 할 과제로 남는다.

첫째, Bounding Box 내 객체 점유율(Bbox-Obj Ratio)에 따른 성능 편차 문제이다. 본 연구의 핵심인 Top-K 위치 추출기는 Bbox 내부에서 활성도가 높은 픽셀이 객체일 것이라는 가정을 전제로 한다. 따라서 MHCD2022 데이터셋의 전차, 장갑차 등과 같이 객체의 형태가 뭉쳐있고(Compact) Bbox 내 점유율이 높은(High-density) 데이터에서는 탁월한 성능을 발휘한다. 그러나 곤충의 더듬이나 가늘고 긴 위장막 지지대처럼 객체가 차지하는 픽셀 비율이 극도로 낮은(Low-density) 경우, Top-K로 추출된 상위 픽셀에 배경(Background)이 다수 포함될 위험이 있다. 이 경우 배경 노이즈가 프로토타입에 섞여 정렬되므로, 탐지 정확도가 저하될 수 있다. 향후 연구에서는 객체의 기하학적 형태를 고려하여, 점유율이 낮은 객체에 대해서도 적응적으로 필터링할 수 있는 정렬 기법이 요구된다.

둘째, Top-K 선택 비율의 수동 설정에 대한 의존성이다.현재 제안하는 방법은 특징 맵에서 추출할 상위 픽셀의 개수 K (혹은 비율 ρ)를 사용자가 실험적으로 결정해야 한다. 이는 데이터셋의 특성이나 입력 해상도가 바뀔 때마다 최적의 K값을 다시 탐색해야 함을 의미하며, 다양한 작전 환경이 존재하는 실제 군사 응용에서는 모델의 범용성을 저해하는 요인이 될 수 있다. 따라서 입력 영상의 복잡도나 객체의 추정 크기에 따라 최적의 K값을 모델이 스스로 학습하여 결정하는 '학습 가능한 Top-K(Learnable Top-K)' 메커니즘 혹은 어텐션 기반의 동적 픽셀 선택 모듈로의 확장이 필요하다.

마지막으로, 본 연구는 초기 백본 네트워크가 추출한 특징의 활성화 값에 의존하여 Top-K 픽셀을 선정한다. 만약 위장 기술이 고도화되어 초기 특징 추출 단계에서 객체가 전혀 활성화되지 않는다면, 이후의 정렬 과정 또한 실패할 가능성이 있다. 이를 보완하기 위해 다른 정보를 활용하거나 의미론적 정보를 보강하기 위한 멀티 모달(Multi-modal) 연구로의 확장이 필요할 것이다.

결론적으로, 본 논문은 MHCD2022와 같은 군사적 데이터셋에 특화된 고효율 위장 객체 탐지 방법론을 제시하였으며, 향후 상기한 한계점들을 보완하여 다양한 전장 환경에서도 강건하게 동작하는 완전 자동화된 무인 감시 시스템의 기반 기술로 발전시킬 수 있을 것이다.

6. 참고 문헌

[1] Redmon, Joseph et al. "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[2] Fan, Deng-Ping et al. "Concealed Object Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6024-6042, 2022.

[3] Lin, Tsung-Yi, et al. "Microsoft COCO: Common Objects in Context." in *Proceedings of European Conference on Computer Vision*, 2014, pp. 740-755.

[4] Fan, Deng-Ping et al. "Camouflaged Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2777-2787, 2020.

[5] Xiao, Fengyang et al. "A Survey of Camouflaged Object Detection And Beyond," in *CAAI Artificial Intelligence Research*, vol. 3, no. 1, pp. 1-28, 2024.

[6] Sun, Jianlin et al. "DRRNet: Macro-Micro Feature Fusion And Dual Reverse Refinement For Camouflaged Object Detection," in *arXiv preprint arXiv:2505.09168*, 2025.

[7] Wang, Kaixin, et al. "PANet: Few-Shot Image Semantic Segmentation with Prototype Alignment." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9197-9206.

[8] Snell, Jake et al. "Prototypical Networks For Few-Shot Learning," in *Advances in Neural Information Processing Systems*, pp. 4077-4087, 2017.

[9] Peng, Yansong et al. "D-FINE: Redefine Regression Task In DETRS as Fine-Grained Distribution Refinement," in *Proceedings of International Conference on Learning Representations*(accepted), 2025.

[10] Liu, Maozhen, Xiaoguang Di, "Extraordinary MHNet: Military high-level camouflage object detection network and dataset," in *Neurocomputing*, pp.126466, 2023.

[11] Lin, Tsung-Yi et al. "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2117-2125, 2017.

[12] Liu, Wei et al. "SSD: Single Shot MultiBox Detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21-37, 2016.

[13] Tan, Mingxing et al. "EfficientDet: Scalable and Efficient Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781-10790, 2020.

[14] Liu, Shu et al. "Path Aggregation Network for Instance Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8759-8768, 2018.

[15] Carion, Nicolas et al. "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision*, pp. 213-229, 2020.

[16] Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965-16974, 2024.

[17] Chen, Qiang et al. "LW-DETR: A Transformer Replacement to YOLO for Real-Time Detection," *arXiv preprint arXiv:2406.03459*, 2024.

[18] Chen, Geng et al. "Camouflaged Object Detection via Context-aware Cross-level Fusion," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6981-6993, 2022.

[19] Pang, Youwei et al. "Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2160-2170, 2022.

[20] He, Shiyu et al. "Cross-Layer Semantic Guidance Network For Camouflaged Object Detection," in *Electronics*, vol. 14, no. 4, pp. 779, 2025.

[21] Z. Cai et al. "Align-DETR: Enhancing End-to-End Object Detection with Aligned Loss," in *British Machine Vision Conference*, 2024.

[22] Wang, Jian et al. "POT: Prototypical Optimal Transport for Weakly Supervised Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15055-15064, 2025.

[23] Kirillov, Alexander et al. "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015-4026, 2023.

[24] Everingham, Mark, et al. "The Pascal Visual Object Classes (VOC) Challenge." in *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[25] Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." in *Advances Neural Information Processing Systems*, 2015, pp. 91-99.