Journal of Institute of Control, Robotics and Systems (2025) 31(5):526-535
https://doi.org/10.5302/J.ICROS.2025.25.0034 ISSN:1976-5622 eISSN:2233-4335

멀티스펙트럴 센서 기반 보행자 검출을 위한 자기 지도 학습 적용 연구

A Study on Application of Self-supervised Learning for Multispectral Pedestrian Detection

허 재 연, 정 의 철, 황 유 진, 최 유 경*
(Jae-Yeon Heo[1], Ui-Cheol Jung[1], Yujin Hwang[1], and Yukyung Choi[1,*])
[1]Department of Artificial Intelligence and Robotics, Sejong University

Abstract: Pedestrian detection has long been a prominent area of focus in computer vision due to its extensive applications, particularly in the development of autonomous driving systems. While color image-based pedestrian detection has achieved significant progress, its vulnerability to illumination changes has prompted more research into multispectral pedestrian detection, which leverages both visible and infrared modalities for robust detection in varying lighting conditions. Despite advances in multispectral detection models, research on applying self-supervised learning for feature representation in this area remains limited. In this study, we extend existing self-supervised learning approaches for color image data to multispectral pedestrian detection. By pre-training with self-supervised learning methods and fine-tuning, we evaluate their extensibility to the multispectral domain. Experimental results demonstrate that self-supervised learning enhances multispectral pedestrian detection, especially in nighttime conditions, outperforming models pre-trained only on color modalities. Furthermore, we analyze the framework of self-supervised learning method that perform well on multispectral data, providing insights into which framework elements enhance multispectral pedestrian detection. This study underscores the potential of self-supervised learning in multispectral pedestrian detection and its role in advancing detection robustness for critical applications such as autonomous driving.

Keywords: multispectral pedestrian detection, representation learning, self-supervised learning

I. 서론

자율 주행 시스템이 안전성을 확보하려면 차량 주변의 보행자를 누락 없이 인식하는 기술이 필수적이다. 이를 위해 딥러닝 기반 객체 검출 기술[1-3]을 활용한 보행자 인식 연구가 활발히 수행되고 있다[4-6]. 하지만 이러한 전통적인 객체 및 보행자 검출 연구는 주로 가시광선 기반의 컬러 영상으로 수행되어 영상 취득 환경의 광량 및 조도 변화에 취약하다는 한계가 있다[7-8].

이러한 한계를 극복하기 위해 야간, 그림자, 악천후 등 저조도 환경에서도 정보 손실이 적은 적외선 열화상 카메라를 함께 활용하는 연구가 주목받고 있다. 열화상 영상과 컬러 영상은 상호 보완적인 정보를 제공하므로, 두 영상 모달리티(modality)를 결합하면 주/야간 조도 변화에 강건한 보행자 인식이 가능하다. 이를 기반으로 다양한 멀티스펙트럴 보행자 검출 연구가 수행되어 보행자 인식의 조도 변화에 따른 안전성이 크게 개선되었다[9-12].

한편, 딥러닝 기반 시스템의 성능 개선을 위해서는 모델 학습에 사용할 양질의 대규모 데이터셋이 필요하다. 그러나, 열화상 영상은 컬러 영상에 비해 공개된 데이터셋이 부족하며, 멀티스펙트럴 데이터셋의 경우 수집 난이도가 높고 영상 간 정렬과 같은 문제로 레이블링(labeling) 및 데이터셋 구축에 많은 비용이 발생한다. 따라서 멀티스펙트럴 보행자 인식 연구에서는 추가 데이터셋 구축 없이 성능을 개선하는 효율적인 방법론이 필요하다.

자기 지도 학습(self-supervised learning)은 데이터에 대한 정답 값(label) 없이 데이터 자체로부터 풍부한 시각적 특징(visual feature)을 학습할 수 있는 방법으로, 레이블링 과정의 필요성을 완화하고 데이터 활용도를 높이는 데 효과적인 대안으로 주목받고 있다. 최근에는 자기 지도 학습 방법이 다양한 벤치마크에서 경쟁력 있는 결과를 보여주어 좋은 시각적 표현을 학습할 수 있는 능력을 입증하였다[13-18]. 하지만 기존의 영상 기반 자기 지도 학습 연구들은 대부분 컬러 영상을 활용하였기 때문에 멀티스펙트럴 데이터에 바로 적용할 수 없었으며, 관련 연구도 아직 활발히 진행되지 않았다[19-20].

* Corresponding Author
Manuscript received February 10, 2025; revised March 10, 2025; accepted April 7, 2025
허재연: 세종대학교 AI로봇학과 대학원생(jyheo@rcv.sejong.ac.kr, ORCID 0009-0000-2254-1105)
정의철: 세종대학교 지능기전공학부 무인이동체공학과 학부생(ucjung@rcv.sejong.ac.kr, ORCID 0009-0009-6638-3044)
황유진: 세종대학교 AI로봇학과 대학원생(yjhwang@rcv.sejong.ac.kr, ORCID 0000-0001-9032-4786)
최유경: 세종대학교 AI로봇학과 부교수(ykchoi@rcv.sejong.ac.kr, ORCID 0000-0002-9970-0132)
※ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 지원(IITP-2025-RS-2022-00156345, 35%)과 과학기술정보통신부의 재원으로 한국연구재단, 무인이동체원천기술개발사업단의 지원(NRF-2023M3C1C1A01098408, 35%)과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2024-00437494, 30%)을 받아 수행되었음.

본 논문은 기존 자기 지도 학습 기반 사전 학습 방법의 멀티스펙트럴 보행자 검출에 대한 확장 가능성을 검토한다. 또한 실험을 통해 자기 지도 학습의 설계 요소 중 어떠한 특성이 멀티스펙트럴 데이터 기반의 사전 학습에 긍정적인 영향을 미치는지 분석하였다. MoCo [14], BYOL [16], SimSiam [17]과 같은 대표적인 컬러 영상 기반 자기 지도 학습 방법론을 적용하였으며, 실세계에 가까운 대표적인 멀티스펙트럴 데이터셋인 KAIST Multispectral Pedestrian Dataset (이하 KAIST 데이터셋)[9]을 활용한 적용 실험에 대한 결과를 분석하였다.

적용 결과, 자기 지도 학습을 멀티스펙트럴 보행자 검출로 확장 가능함을 확인하였으며, 영상의 국소적 정보(local information)를 활용한 자기 지도 학습법이 효과적이라는 사실 및 열화상 영상의 특성을 충분히 활용하기 위해서는 기존 방법론에 개선이 필요하다는 점을 확인하였다. 추가적으로, 기존 자기 지도 학습 방법론 중 적용 결과 가장 높은 학습 성능을 보인 SoCo [21] 프레임워크에 기반한 분석을 수행하여 멀티스펙트럴 보행자 검출에 자기 지도 학습 적용 시 고려해야 할 요소를 제시하였다. 분석 결과, 검출기의 두 입력 인코더(encoder)를 동일한 가중치로 초기화하는 기존 방법보다 입력 데이터의 특성을 고려해 설계된 가중치로 각 인코더를 별도로 초기화하는 방법이 효과적임을 확인했으며, 사전 학습 단계에서 영상의 국소적 정보를 충분히 학습할 수 있도록 객체 수준 표현(object-level representation)을 학습에 활용하는 것이 중요함을 확인하였다. 또한 사전 학습 단계에서 컬러 영상만을 활용했을 때보다 열화상 영상을 함께 활용했을 때 저조도 상황에서 일관적으로 검출 성능을 개선할 수 있음을 확인하였다.

본 논문의 이후 구성은 다음과 같다. 제2장에서는 수행 연구와 관련된 딥러닝 기반 멀티스펙트럴 보행자 검출 및 자기 지도 학습의 연구 동향에 대해 서술한다. 제3장에서는 멀티스펙트럴 보행자 검출을 위한 자기 지도 학습법 설계의 필요성을 검토한다. 제4장에서는 다양한 실험 결과를 기반으로 멀티스펙트럴 보행자 검출을 위한 자기 지도 학습 설계 시 고려 요소를 자세히 분석하고, 효과적인 적용 기법을 제안한다. 5장에는 3, 4장에서 수행된 실험들의 세부적인 구현 사항을 정리한다. 마지막으로 제6장에는 결론을 서술한다.

II. 관련 연구

1. 멀티스펙트럴 보행자 검출

멀티스펙트럴 보행자 검출은 보행자 인식의 정확도와 안정성을 높이기 위해 다중 센서 기반으로 객체 인식을 수행하는 연구 분야로, 특히 저조도 환경이나 기상 조건이 열악한 상황에서도 높은 성능을 보장하기 위해 고안되었다[9]. 해당 연구는 일반적으로 가시광선 영역의 컬러 카메라 영상을 적외선 영역의 열화상 카메라 영상과 결합하여 안정적인 보행자 인식을 수행하는 것을 목적으로 한다. 특히 열화상 카메라는 조도 변화에 강건하게 환경 정보를 습득할 수 있기 때문에 저조도 환경에서의 보행자 검출 성능을 향상시키기 위한 핵심적인 센서로 주목받았으며, 이에 열화상 영상을 함께 활용하여 보행자를 인식하려는 연구가 활발히 진행되어 조도 변화에 따른 보행자 인식률을 크게 개선할 수 있었다[10-12].

열화상 영상을 활용한 멀티스펙트럴 보행자 검출 연구의 대표적인 사례로 MLPD [10]가 있다. 해당 연구는 SSD [2] 네트워크를 기반으로 컬러 영상 및 열화상 영상의 시각적 특징을 융합하는 기법을 도입하여 주/야간 다양한 조명 조건에서 강건한 검출 성능을 달성하였다. 그러나 이러한 기존 연구들은 주로 검출기 내부에서 멀티스펙트럴 데이터 간 특징 융합 방법을 개선하는 데 초점을 맞추는 경향이 있었으며[10,11,12,22], 멀티스펙트럴 데이터셋의 활용 효율성을 높여 인식 성능을 개선하기 위한 시도는 상대적으로 적었다.

멀티스펙트럴 보행자 검출 모델의 학습에 활용되는 데이터는 수집 과정에서 높은 비용이 발생할 뿐만 아니라 취득 영상 간 정렬과 같은 기술적 문제를 수반하기 때문에 데이터셋 활용도를 높이는 방안이 중요한 과제이다. 본 연구는 기존의 자기 지도 학습 방법을 멀티스펙트럴 보행자 검출에 적용하여 사전 학습 과정에서 검출 성능에 영향을 미치는 요소를 분석하고, 멀티스펙트럴 데이터 활용의 효율성을 높이기 위한 방안을 체계적으로 검토하였다. 이를 통해 추가적인 데이터셋 구축 비용 없이 멀티스펙트럴 보행자 검출 모델의 인식 성능 개선을 위한 방법을 제시한다.

2. 자기 지도 학습

자기 지도 학습은 데이터에 대한 레이블(label) 없이 모델이 데이터 자체의 표현(representation) 및 특징(feature)을 학습하는 방법으로 모델의 사전 학습에 주로 사용되는 기법이다. 사전 학습을 통해 얻어진 가중치는 모델을 초기화하는데 사용되며, 이후 모델은 다운스트림 작업에 알맞게 미세 조정(fine-tuning)된다. 최근에는 데이터 증강(data augmentation)을 통해 생성된 다양한 변형 영상 간의 임베딩 공간 내부 거리를 조정하는 대조 학습(contrastive learning)을 기반으로 데이터에 대한 표현력을 학습하는 연구[14-17]들이 대용량의 데이터를 활용하여 지도학습에 비견될 정도의 학습 성능을 보이고 있다. 또한, 객체 검출(object detection)이나 의미론적 분할(semantic segmentation)과 같이 영상의 국소적 정보(local information)를 필요로 하는 작업을 위해 영상의 부분적인 특징을 더욱 효과적으로 학습할 수 있는 기법들이 제안되었다[23-26]. 이들 연구는 기존 대조 학습 방법론에서 생성되는 특징 벡터가 영상의 국소 부분에 대한 정보를 더 풍부하게 포함할 수 있도록 프레임워크를 설계하였으며, 이를 통해 객체 검출 및 의미론적 분할과 같은 작업에서 성능을 효과적으로 개선하였다.

그러나 멀티스펙트럴 데이터 기반 객체 검출 분야에서는 자기 지도 학습의 적용 및 최적 설계와 관련된 연구가 상대적으로 미진한 상황이다. 특히, 열화상 영상과 컬러 영상의 상호 보완적 특성을 활용하는 멀티스펙트럴 데이터 환경에서 기존 방법론이 이러한 데이터의 특성을 충분히 반영하지 못하고 있다.

본 연구는 기존의 자기 지도 학습 방법론을 멀티스펙트럴 보행자 검출 프레임워크에 확장하고, 해당 환경에서 효과적인 자기 지도 학습법 설계를 제안하여 그 효과를 실험적으로 검증하였다. 이를 통해 멀티스펙트럴 데이터의 활용도를 높

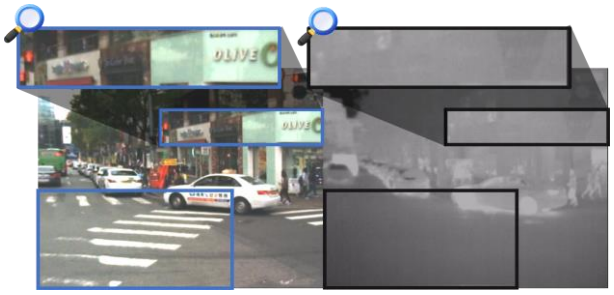

그림 1. 동일 주행 장면에 대한 모달리티 별 정보 비교 도식. (왼쪽) 컬러 영상, (오른쪽) 열화상 영상. 열화상 영상에는 사물의 색상, 질감과 같은 맥락적 정보가 부족하다.

Fig. 1. Comparison of modality-specific information for the same driving scene. (Left) color image, (Right) thermal image. The thermal image lacks contextual information such as the color and texture of objects.

이고, 자기 지도 학습 환경에서 보행자 검출 성능을 개선하기 위한 방법을 제시하고자 한다.

III. 멀티스펙트럴 보행자 검출을 위한 자기 지도 학습법 설계 필요성 검토

1. 멀티스펙트럴 보행자 검출로의 자기 지도 학습 적용 및 그 한계

자기 지도 학습을 사전 학습(pre-training)에 활용할 때 중요한 설계 요소 중 하나는 목표로 하는 다운스트림 작업(downstream task)의 특성을 고려하는 것이다. 컴퓨터 비전 분야에서 다운스트림 작업은 영상의 전역 정보(global information)를 활용하는 과제와 국소 정보(local information)를 요구하는 과제로 나뉜다.

보행자 검출은 입력 영상에서 보행자의 위치를 정확히 탐지하는 작업으로 영상의 국소 정보에 의존하는 과제에 속한다. 그러나 기존 연구에서 제안된 국소 정보 기반 자기 지도 학습 방법론이 멀티스펙트럴 보행자 검출에 적합하다고 단정할 수는 없다. 이는 기존 자기 지도 학습 연구에서 열화상 영상을 입력으로 사용하는 경우가 충분히 고려되지 않았기 때문이다. 그림 1은 두 영상 모달리티의 차이를 시각화 하기 위해 동일 주행 장면에 대한 컬러 카메라 영상(왼쪽)과 열화상 카메라 영상(오른쪽)을 나타낸 것이다. 그림 1에서 컬러 영상과 비교해 열화상 영상에는 횡단보도 및 차선과 같은 교통노면표시, 건물의 간판과 같은 정보가 부재하다. 자기 지도 학습을 적용할 경우 컬러 영상은 풍부한 텍스처와 색상 정보를 기반으로 학습이 가능한 반면, 열화상 영상은 비교적 맥락적 정보(contextual information)가 부족하고 주로 온도 분포에 기반한 단조로운 윤곽 정보만을 포함하고 있어 자기 지도 학습의 효과가 제한될 가능성이 있다. 따라서 멀티스펙트럴 보행자 검출에 기존의 자기 지도 학습 방법론이 잘 동작하는지 검증이 필요하다.

본 연구에서는 전역 정보 기반의 자기 지도 학습 방법과 국소 정보 기반의 자기 지도 학습 방법을 멀티스펙트럴 보행자 검출 모델의 사전 학습 설계에 각각 적용하여 그 결과와 영향을 분석하였다. 실험 결과(표 1 참조), DetCo [25], CCrop [26], SoCo [21]와 같은 국소 정보(local 특징)에 초점을 맞춘 학습 방법론이 MoCo [14], SimSiam [17], BYOL [16]과 같은 전역 정보(global 특징)를 활용하는 방법론 대비 학습 효과가 더 우수하다는 점을 확인하였다. 그러나 열화상 영상에 대한 의존성이 큰 야간 환경에서는 검출 성능 개선이 제한되는 현상이 관찰되었다. 이를 통해 열화상 영상의 특성을 반영하여 기존 방법론을 개선할 필요가 있음을 확인할 수 있었다.

표 1. 자기 지도 학습 방법론 확장 적용 결과. Miss Rate 값이 낮을수록 더 우수한 검출 성능을 의미한다.

Table 1. Results of self-supervised learning method applications. Lower miss rate indicates better detection performance.

Method	사전 학습 활용 특징	Miss Rate(IoU = 0.5) ↓		
		All	Day	Night
Xavier	-	14.46	16.70	9.85
MoCo	Global	13.92	15.25	11.18
SimSiam	Global	14.18	15.52	11.01
BYOL	Global	14.46	15.07	12.93
DetCo	Local	12.31	12.30	12.14
CCrop	Local	11.81	11.43	12.25
SoCo	Local	11.79	12.24	10.84

2. 멀티스펙트럴 보행자 검출에 대한 기존 자기 지도 학습 적용의 한계 분석

멀티스펙트럴 보행자 검출을 위한 기존 자기 지도 학습 방법론의 확장 적용 결과는 표 1에 요약되어 있다. 본 실험에서는 각 자기 지도 학습 방법으로 사전 학습을 진행하고, 이후 사전 학습 가중치로 멀티스펙트럴 검출기를 초기화한 다음 미세 조정하여 검출 성능을 평가하였다. 이 때 모달리티 별 학습의 적합성을 분석하기 위해 주/야간 환경에서 개별적으로 성능을 평가하였다. 이는 주간 환경에서 컬러 영상 기반의 인식 성능과 야간 환경에서 열화상 영상 기반의 인식 성능을 확인하기 위함이다. 성능 평가는 Miss Rate [4]을 주요 지표로 사용하였으며, 값이 낮을수록 검출 성능이 우수함을 의미한다. 조도 변화에 따른 검출률을 비교하기 위해 종일(All), 주간(Day), 야간(Night)에 대한 Miss Rate을 표 1에 별도로 나타내었다. 해당 실험의 추가적인 세부 설계 사항은 5장에 정리하였다.

실험 결과, 모든 자기 지도 학습 방법론이 Xavier [27] 초기화 방식보다 Miss Rate (All)에서 낮은 값을 기록하여 사전 학습을 활용하지 않는 것보다 효과적임을 확인하였다. 특히, 국소 정보 기반 학습에 초점을 맞춘 방법론들(DetCo [25], CCrop [26], SoCo [21])은 전역 정보 기반 학습 방법론들(MoCo [14], BYOL [16], SimSiam [17])보다 우수한 성능을 나타내었다. 이는 멀티스펙트럴 보행자 검출 작업의 특성을 고려하였을 때, 자기 지도 학습을 활용한 사전 학습 과정에 국소 정보에 대한 표현력 개선을 고려한 설계가 필요함을 의미한다.

세부적으로, SoCo는 DetCo와 CCrop 대비 낮은 Miss Rate (All)인 11.79를 기록하며 가장 우수한 성능을 보였다. 특히 SoCo는 야간 환경에서도 10.84의 가장 낮은 Miss Rate (Night)를 기록하여 타 방법론들보다 비교적 열화상 영상 데이터의 특성을 반영할 수 있는 학습법임을 알 수 있었다. 그러나

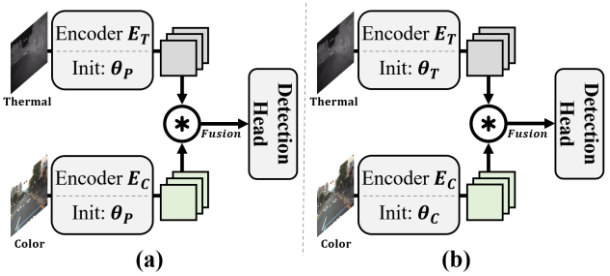

그림 2. 멀티스펙트럴 검출기의 가중치 초기화 방법 비교 도식. (a) 기존의 초기화 방법. (b) 제안하는 초기화 방법. (a)는 검출기의 두 입력 인코더 E_T, E_C를 동일한 특정 사전 학습 가중치 θ_P로 초기화한다. (b)는 E_T를 열화상 영상으로 대조 학습한 가중치 θ_T로, E_C를 컬러 영상으로 대조 학습한 가중치 θ_C로 초기화한다.

Fig. 2. Comparison of weight initialization methods for the multispectral detector. (a) Conventional initialization method, (b) Proposed initialization method. In (a), both encoders E_T and E_C are initialized with the same pretrained weight θ_P. In (b), encoder E_T is initialized with contrastively learned weights θ_T from thermal images, while encoder E_C is initialized with contrastively learned weights θ_C from visible images.

모든 방법론이 야간 환경에서는 Xavier 초기화보다 높은 Miss Rate (Night)를 보였다. 이를 통해 열화상 영상의 맥락 정보 부족으로 인하여 자기지도 학습 방법을 활용한 성능 개선에 한계가 있음을 확인할 수 있다.

정리하면, 멀티스펙트럴 보행자 검출을 위한 자기 지도 학습을 설계하기 위해서는 국소 정보를 학습에 활용해야 함을 알 수 있었다. 또한, 단순히 컬러 영상을 함께 활용해 표현력을 학습한 가중치로는 열화상 영상 기반의 인식 성능 개선에 한계가 있어 각 모달리티에 특화된 학습 구조를 설계해야 함을 확인하였다.

IV. 멀티스펙트럴 보행자 인식을 위한 자기 지도 학습 설계 시 고려 요소

1. 멀티스펙트럴 보행자 인식을 위한 자기 지도 학습 개선 방법

3장 표 1의 실험을 통해서 기존 자기 지도 학습 기반의 가중치 초기화 방법이 열화상 도메인을 포함한 멀티스펙트럴 영상 기반 보행자 인식 성능 개선에 한계를 가진다는 것을 확인하였다. 우리는 자기 지도 학습을 통한 사전 학습이 멀티스펙트럴 보행자 검출로의 적용에 적합하도록 개선하기 위해 단일 가중치를 사용하는 대신 모달리티 특화 가중치를 사용하도록 초기화 방식을 변경하였다. 그림 2에는 멀티스펙트럴 검출기의 기존 초기화 방법과 제안하는 초기화 방법을 비교해 나타내었다.

멀티스펙트럴 보행자 인식 모델은 일반적으로 별도의 모달리티별 영상을 입력 받는 별개의 인코더(E_T, E_C)를 포함한다. 열화상 및 컬러 영상은 각각 인코더 E_T, E_C에 입력되어 특징 맵으로 임베딩되며, 이후 상호 보완적인 정보를 갖는 열화상 영상 특징맵과 컬러 영상 특징맵이 모델 중간에서 융합(fusion)되어 최종 예측에 사용된다. 기존 방법[10-12]의 경우,

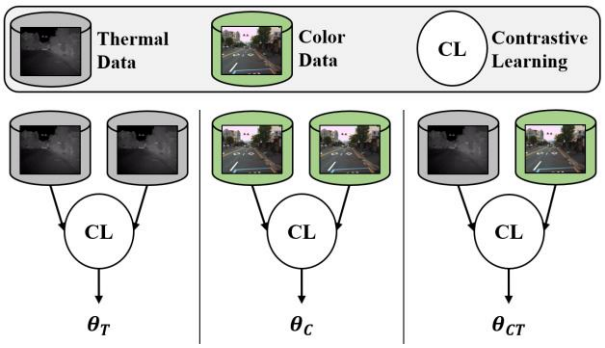

그림 3. 멀티스펙트럴 데이터를 활용한 사전 학습 가중치 생성 도식. θ_T는 열화상 영상으로 대조 학습한 가중치, θ_C는 컬러 영상으로 대조 학습한 가중치, θ_{CT}는 컬러와 열화상 영상을 함께 학습한 가중치이다.

Fig. 3. Diagram of generating pretrained weights using multispectral data. θ_T represents the weights trained on thermal images via contrastive learning, θ_C represents the weights trained on color images via contrastive learning, and θ_{CT} represents the weights jointly trained on both color and thermal images.

그림 2의 (a)와 같이 두 인코더 E_T, E_C를 모두 동일한 특정 가중치 θ_P로 초기화했기에 각 모달리티 특성을 고려한 가중치를 활용하지 못하였다.

이에 우리는 입력 영상의 모달리티 특성을 고려한 가중치를 활용하기 위해 그림 2의 (b)와 같은 별도 초기화 방식을 제안하였다. 해당 방법은 열화상 영상으로 사전 학습한 가중치 θ_T로 인코더 E_T를 초기화하고, 컬러 영상으로 사전 학습한 가중치 θ_C로 인코더 E_C를 초기화하도록 구성하였다.

추가적으로, 우리는 SoCo 프레임워크를 기반으로 열화상 영상에 대한 표현력 학습을 개선한 가중치를 설계했을 때의 보행자 인식 성능 개선 효과를 검토하였다. 표 2의 실험에서, 열화상 도메인에 최적화된 사전 학습 가중치를 멀티스펙트럴 검출기 초기화에 함께 사용했을 때 성능이 Xavier [27] 초기화 방식의 결과보다 전반적으로 크게 개선되어 보행자 인식률이 향상될 수 있음을 확인하였으며, 표 3의 추가적인 실험을 통해 해당 경향성이 일관적으로 유지됨을 확인하였다. 표 4, 표 5의 실험을 통해 기존의 컬러 영상 기반 학습에서 연구된 국소 영역을 활용하는 사전 학습 방법이 열화상 도메인에서도 효과적으로 확장될 수 있음을 확인했으며, 표 6의 실험을 통해 두 영상 모달리티 각각의 특수성을 반영하는 사전 학습 가중치를 설계하는 것이 중요함을 확인하였다.

본 연구는 데이터셋 추가 구축 비용 없이 멀티스펙트럴 보행자 인식의 성능을 효과적으로 개선하기 위한 자기 지도 학습 설계 요소를 제시하였다. 특히 열화상 도메인 특성을 고려한 설계로 기존 자기 지도 학습의 한계를 개선하였다.

2. 모달리티별 가중치 초기화 적용 결과 분석

멀티스펙트럴 검출기의 입력 인코더 초기화 전략에 따른 성능 비교 결과는 표 2와 같다. 해당 실험에서는 멀티스펙트럴 검출기의 인코더 E_T와 E_C를 초기화하기 위해 그림 3과 같이 세 가지 유형의 가중치 θ_T, θ_c, θ_{CT}를 설계하였다. θ_T는 열화상 도메인 데이터로 대조 학습한 가중치, θ_C는 컬러 도메인 데이터로 대조 학습한 가중치이며, θ_{CT}는 컬러와 열

표 2. 모달리티 특성을 고려한 멀티스펙트럴 검출기 인코더 초기화 전략 비교 실험 결과.

Table 2. Experimental results of comparing encoder initialization strategies for multispectral object detection based on modality-specific features.

초기화 방안	인코더 E_T	인코더 E_C	Miss Rate(IoU = 0.5) ↓		
			All	Day	Night
단일 가중치	θ_{CT}	θ_{CT}	12.08	13.55	8.70
	θ_T	θ_T	13.49	14.93	9.61
	θ_C	θ_C	12.06	12.03	11.60
모달리티 고려	θ_T	θ_C	11.39	12.34	8.76

표 3. 다양한 백본 네트워크에서의 검증 실험 결과.

Table 3. Experimental results from validation across different backbone networks.

backbone	인코더 초기화 가중치	Miss Rate(IoU = 0.5) ↓		
		All	Day	Night
ResNeXt50	θ_{CT}	14.24	15.53	10.73
	θ_T / θ_C	13.49	14.48	10.27
ResNet50(×2)	θ_{CT}	14.68	16.66	10.80
	θ_T / θ_C	12.65	13.34	9.53
ResNet101	θ_{CT}	15.63	16.51	13.94
	θ_T / θ_C	15.42	15.69	13.30

화상 도메인 데이터를 대조 학습에 함께 사용하여 학습한 가중치이다, 각 가중치를 활용해 인코더 E_T와 E_C를 초기화하였으며, 각각의 초기화 방법이 보행자 검출 성능에 미치는 영향을 평가하였다.

표 2의 결과에서 E_T와 E_C를 동일한 가중치로 초기화한 경우, θ_C로 초기화한 모델이 Miss Rate (All) 12.06로 비교적 우수한 성능을 보였지만 저조도 환경에서 Miss Rate (Night) 11.60로 성능이 크게 저하되었다. 반면, θ_{CT}만을 활용한 경우 저조도 환경에서 Miss Rate (Night) 8.70으로 높은 개선을 보였으나 주간 환경에서의 성능 개선은 제한적이었다. 이는 사전 학습 과정에서 컬러 모달리티의 색상 및 질감, 열화상 모달리티의 윤곽과 같은 상호 보완적인 정보를 충분히 학습하지 못했기 때문으로 해석된다. θ_T만을 활용한 경우에는 주/야간 모두에서 비교적 낮은 성능을 기록했으며, 이는 열화상 영상만을 사전 학습에 사용하는 것이 멀티스펙트럴 보행자 검출에 적합하지 않음을 시사한다.

한편, 모달리티 특성을 고려하여 각 인코더 E_T, E_C를 입력 모달리티에 특화된 가중치 θ_T, θ_C로 초기화한 방식은 모든 초기화 방안 중 가장 높은 성능을 나타냈다. 이 방식은 Miss Rate (All)에서 11.39를 기록하였으며, 동시에 저조도 환경에서도 Miss Rate (Night) 8.76로 우수한 결과를 보였다. 이는 입력 데이터의 모달리티 특성을 고려하여 초기화한 방식이 멀티스펙트럴 보행자 검출 성능 개선에 적합함을 의미한다. 이러한 경향성이 일관적으로 유지되는지 검증하기 위해 우리는 다양한 백본 인코더를 활용해 추가적으로 실험을 수행하였으며, 그 결과를 표 3에 정리하였다. 표 3의 실험에서는 이전 표 1, 2의 실험에서 검출기의 인코더 네트워크로 사용되었던 ResNet50 [28] 백본을 ResNeXt50 [29], ResNet50(×2) [30], ResNet101 [28]으로 교체하여 가중치 초기화 전략의 영향을 추가적으로 평가하였다. 표 3의 '인코더 초기화 가중치'에서 θ_{CT}은 검출기의 두 입력 인코더를 모두 동일 사전 학습 가중치 θ_{CT}로 초기화 시킨 것을 의미하며, θ_T / θ_C는 검출기의 인코더 E_T를 가중치 θ_T로, 인코더 E_C를 가중치 θ_C로 초기화하여 평가한 것을 의미한다. 세 종류의 다른 인코더를 활용한 실험 결과, 각 인코더 E_T, E_C를 입력 모달리티에 특화된 가중치 θ_T, θ_C로 초기화하는 전략이 단일 가중치로 초기화하는 것보다 일관적으로 개선된 결과를 보임을 확인하였다.

표 4. 영역 정보 추출 방법 실험 결과.

Table 4. Experimental results of region information extraction methods.

영역 추출 방법	Miss Rate(IoU = 0.5) ↓		
	All	Day	Night
Random Box	11.96	12.32	10.87
Patch	11.74	12.63	9.55
Selective Search	11.39	12.34	8.76

표 5. 객체 수준 특징 추출 실험 결과.

Table 5. Experimental results of object-level feature extraction.

FPN 및 R-CNN head 활용 여부	Miss Rate(IoU = 0.5) ↓		
	All	Day	Night
X	11.86	12.89	9.64
O	11.39	12.34	8.76

결론적으로, 멀티스펙트럴 보행자 검출 성능을 극대화하기 위해서는 각 모달리티에 특화된 가중치를 인코더 초기화에 사용하는 것이 효과적임을 확인하였다. 또한, 단일 모달리티 기반 사전 학습 방식보다는 컬러 영상과 열화상 영상을 함께 활용한 사전 학습 방식이 주/야간 환경 변화에 강건한 성능을 제공한다는 점을 입증하였다.

3. 열화상 도메인을 위한 사전 학습 가중치 설계 개선 결과 분석

열화상 도메인에서 SoCo [21]의 국소 지역 정보 표현력 개선을 위한 모듈의 확장성을 분석한 결과는 표 4와 표 5에 제시되어 있다. SoCo는 객체 단위의 표현력을 학습하기 위해 대조 학습 쌍(pair) 구성 시 비지도 학습 기반 객체 영역 제안 기술인 Selective Search [31]를 활용한다. 본 연구에서는 멀티스펙트럴 데이터 환경에서 Selective Search의 유효성을 평가하기 위해 Selective Search로 생성한 영역, 무작위로 생성한 박스(random box) 영역, 이미지를 나누어 생성한 패치(patch) 영역 각각을 기반으로 대조 학습 쌍을 구성하여 사전 학습 효과를 비교하였다.

표 4에 나타난 실험 결과, Selective Search로 생성된 영역을 활용한 대조 학습이 11.39의 가장 낮은 Miss Rate (All)을 기록하여 패치나 랜덤 박스를 사용하는 경우보다 학습에 더 효과적임을 확인하였다. 특히 저조도 야간 환경에서 Miss Rate (Night) 8.76을 기록해 Selective Search의 이점이 두드러졌으며, 이는 열화상 도메인의 컬러 영상 대비 적은 맥락(context)으로 인한 표현력 학습의 한계를 적절한 영역 쌍 생성 방식을 통해 보완했기 때문으로 분석된다. 그림 4는 동일한 주행 장면의 열화상 영상과 컬러 영상 각각에서 생성된 Selective Search 제안 박스 P_T, P_C를 시각화한 결과이다. 그림 4를 통해 Selective Search가 열화상 영상에서 정보량이 적은 영역을 효과적으로 배제하고, 정보량이 많으며 주행 장면 이 해에

그림 4. Selective Search 적용 결과 시각화. (왼쪽) 컬러 영상 및 제안 영역 P_C, (오른쪽) 열화상 영상 및 제안 영역 P_T. 제안된 영역 P_C, P_T는 빨간 박스로 나타내었다.

Fig. 4. Visualization of selective search results. (Left) color images and proposed regions P_C, (Right) thermal images and proposed regions P_T. The proposed regions P_C, P_T are highlighted with red boxes.

중요한 보행자, 차량, 도로 환경 구조물과 같은 영역을 효과적으로 제안할 수 있음을 확인하였다.

다음으로, FPN (Feature Pyramid Network)과 R-CNN Head를 도입하여 객체 수준 표현력을 개선하는 설계를 자기 지도 학습에 포함했을 때의 사전 학습 가중치 효과를 멀티스펙트럴 보행자 검출에서 평가하였다. 실험 결과는 표 5에 제시되어 있으며, FPN과 R-CNN Head를 사전 학습 구조에 통합한 경우 Miss Rate (All), Miss Rate (Day), Miss Rate (Night) 모든 조건에서 더 우수한 검출 성능을 보였다. 이는 사전 학습 단계에서 영상의 국소적인 정보를 포함하는 객체 수준 표현을 효과적으로 학습하면 다운스트림 작업인 멀티스펙트럴 보행자 검출 성능이 유의미하게 향상됨을 시사한다.

추가적으로, 영역 제안 방식이 열화상 도메인에서 사전 학습 성능 개선에 미치는 영향을 분석하기 위해 제안 박스 P_T, P_C를 활용하여 대조 학습 쌍을 구성하였다. 이러한 박스를 다양한 방식으로 조합하여 사전 학습을 진행해 검출기의 각 인코더 E_T, E_C 초기화를 위한 가중치를 설계하였다. 표 6에는 각 인코더 E_T, E_C를 초기화하기 위한 사전 학습 가중치 설계 방안 및 그에 따른 검출 결과를 나타내었다.

표 6의 결과에 따르면, E_T 가중치의 사전 학습을 설계할 때는 열화상 영상 및 P_T만을, E_C 가중치의 사전 학습을 설계할 때는 컬러 영상 및 P_C만을 개별적으로 활용하는 방식이 종일, 주간, 야간 환경에서 전반적으로 가장 좋은 검출 결과를 보였다(표 6의 가장 아래 행). 하지만 P_T와 P_C를 모두 사용하는 경우($P_T + P_C$), 학습에 사용되는 대조 쌍의 수가 증가함에 따라 각 모달리티에 적합하지 않은 노이즈 영역이 추가되어 위 방식과 비교해 학습에 부정적인 영향을 미쳤다.

반면, 사전 학습 시 단일 모달리티(열화상 또는 컬러 영상)만 활용하는 경우에는 각 모달리티의 제안 영역을 합친

표 6. 모달리티 및 제안 박스 활용에 대한 실험 결과. P_T는 열화상 영상의 제안 영역, P_C는 컬러 영상의 제안 영역을 의미한다.

Table 6. Experimental results on modality and region proposal box utilization. P_T denotes the region proposal from thermal images, and P_C denotes the region proposal from color images.

E_T 가중치 설계		E_C 가중치 설계		Miss Rate(IoU = 0.5) ↓		
제안 영역	학습 데이터	제안 영역	학습 데이터	All	Day	Night
$P_T + P_C$	Thermal	$P_T + P_C$	Color	12.23	12.60	10.38
$P_T + P_C$	Color	$P_T + P_C$	Color	11.26	11.03	11.26
$P_T + P_C$	Thermal	$P_T + P_C$	Thermal	11.71	12.39	10.02
P_T	Thermal	P_T	Color	11.81	12.71	9.29
P_T	Thermal	P_T	Thermal	13.49	14.93	9.61
P_C	Thermal	P_C	Color	12.54	13.47	10.77
P_C	Color	P_C	Color	12.06	12.03	11.60
P_T	Thermal	P_C	Color	11.39	12.34	8.76

$P_T + P_C$를 모두 대조 학습 쌍으로 사용하는 것이 P_T, P_C를 개별적으로 활용하는 것보다 효과적이었다. 컬러 데이터로 학습한 경우 P_C보다 $P_T + P_C$를 활용했을 때, 열화상 데이터로 학습한 경우 P_T보다 $P_T + P_C$를 활용했을 때에 낮은 Miss Rate (All)를 기록하였다. 이는 사전 학습에 단일 모달리티 영상 데이터만 사용할 경우 해당 모달리티에서 비교적 강조되지 않는 특징 정보를 추가적인 영역 제안을 통해 보완적으로 학습할 수 있음을 의미한다. 이러한 결과는 멀티스펙트럴 보행자 검출을 위한 사전 학습 설계 시 각 모달리티의 특수성을 최대한 반영하면서도, 단일 모달리티 환경에서는 보완적인 학습 방법론을 적용하는 것이 중요함을 의미한다.

마지막으로, E_T 및 E_C 초기화 가중치를 사전 학습할 때 컬러 영상만을 사용할 경우 Miss Rate (Night) 성능이 저조하였지만, 열화상 영상을 함께 활용할 경우 저조도 환경의 Miss Rate (Night) 성능이 일관적으로 개선되어 멀티스펙트럴 검출을 위한 자기 지도 학습 시 두 도메인 데이터 모두를 활용하는 이점을 다시 한번 확인하였다.

V. 각 실험의 세부 구현 방안

1. 베이스라인 프레임워크

각 실험에서는 자기 지도 학습 방법론으로 사전 학습을 수행해 가중치를 얻은 다음, 사전 학습 가중치로 멀티스펙트럴 검출기를 초기화한 뒤 미세 조정하여 검출 성능을 평가하였다. 표 2, 3, 4, 5, 6의 실험은 표 1의 실험에서 가장 좋은 결과를 보여준 SoCo [21] 프레임워크를 기반으로 수행되었다. SoCo는 객체 수준 표현 간의 유사성을 최대화하는 대조 학습을 통해 객체 탐지 작업에 적합한 사전 학습을 수행하는 프레임워크이다. 전체적인 사전 학습 과정은 그림 5와 같다.

SoCo는 객체 수준 표현을 학습에 활용하기 위해 비지도 학습 기반 영역 제안(region proposal) 방법인 Selective Search [31] 및 샘플링(random sampling)을 통해 대조 학습의 기본 연산 단위인 객체 제안 영역을 생성한 뒤, 제안된 영역 및

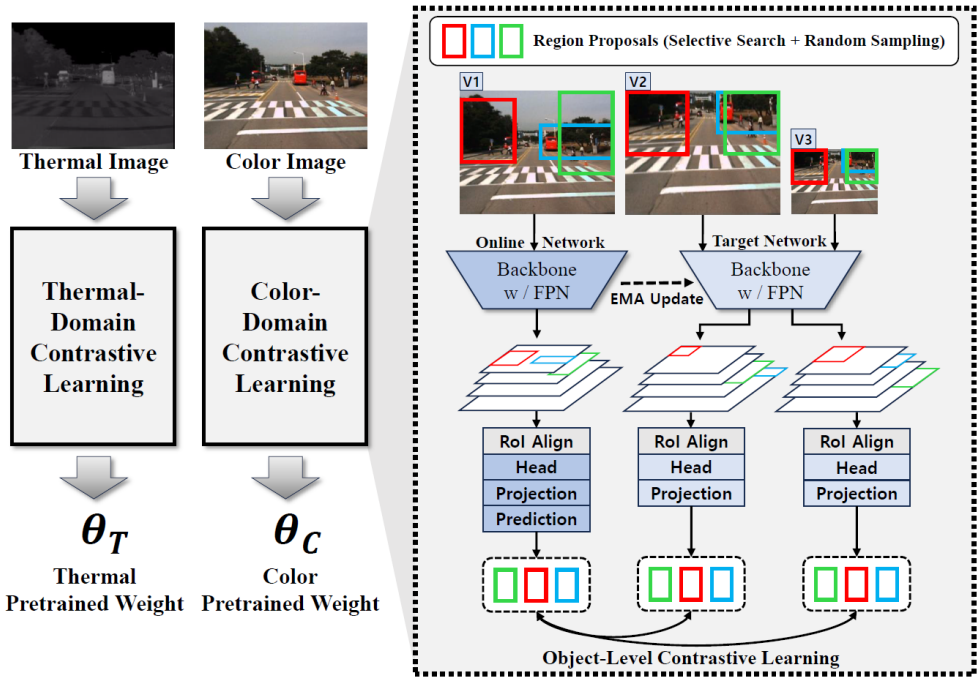

그림 5. 실험에 사용된 사전 학습 구조도. 열화상 영상과 컬러 영상 각각에 SoCo 사전 학습을 적용하여 사전 학습 가중치(θ_T, θ_C)를 생성한다. 각 가중치는 멀티스펙트럴 검출기의 두 인코더를 초기화하는데 사용된다. 사전 학습 과정에서 각 영상에 Selective Search를 적용하여 객체 제안 영역을 얻은 다음, 동일 객체의 크기와 위치가 다른 3개의 뷰 $\{V1, V2, V3\}$를 만든다. 각 뷰를 백본에 통과시켜 영상 단위 특징을 만든 후, RoIAlign을 통해 객체 수준 특징을 만든다. 객체 수준 특징을 활용하여 대조 학습으로 Online Network를 학습하며, Target Network는 exponential moving average(EMA)를 통해 업데이트 된다. 각 실험에서는 사전 학습을 통해 얻어진 가중치로 멀티스펙트럴 검출기의 각 인코더를 초기화한 다음, 검출기를 미세 조정하여 평가를 진행하였다.

Fig. 5. Pretraining framework used in the experiments. SoCo pretraining is applied separately to thermal and color images to generate pretrained weights(θ_T, θ_C). These weights are used to initialize the two encoders of the multispectral detector. During pretraining, Selective Search is applied to each image to obtain object region proposals, followed by the construction of three views $\{V1, V2, V3\}$ with different scales and locations of the same object. A backbone with FPN is used to extract image-level features, and RoIAlign is applied to obtain object-level features. Contrastive learning is then used to train the Online Network, while the Target Network is updated via an exponential moving average (EMA). In each experiment, the pretrained weights are used to initialize the encoders of the multispectral detector, which is then fine-tuned for evaluation.

영상을 서로 다르게 증강하여 세 개의 뷰 $V1$, $V2$, $V3$를 생성한다. 이 때 $V2$는 $V1$에 crop 및 resize를 적용하여 생성하고, $V3$는 $V2$를 downsampling하여 생성한다. 이후 증강된 각 뷰는 FPN (Feature Pyramid Network)이 포함된 인코더를 거쳐 특징 맵으로 임베딩 되고, RoIAlign과 R-CNN Head를 적용하여 객체 수준 표현 h를 추출한다. 이 과정은 아래의 수식 (1)로 표현된다.

$$h = f^H\left(RoIAlign(f^I(V), b)\right) \quad (1)$$

여기서 f^I는 FPN이 포함된 인코더를, f^H는 R-CNN Head를 나타내며, V는 뷰, b는 박스 정보를 의미한다.

SoCo의 학습 과정에는 Online Network, Target Network가 사용된다. Online Network는 입력 뷰 $V1$에서 임베딩 벡터 v_i를 생성하고, Target Network는 입력 뷰 $V2$와 $V3$에서 각각 임베딩 벡터 v'_i, v''_i를 생성한다. Online Network는 R-CNN Head 이후 Projector g_θ와 Predictor q_θ를 거치도록 하며, Target Network는 R-CNN Head 이후 Projector g_ξ만을 거쳐 대조 학습에 필요한 객체 수준 임베딩 벡터를 생성한다. 이 과정은 아래 수식 (2)와 같이 표현할 수 있다.

$$v_i = q_\theta\left(g_\theta(h_i)\right), v'_i = g_\xi(h'_i), v''_i = g_\xi(h''_i) \quad (2)$$

대조 학습은 각 제안 영역에 대해 임베딩 벡터 간 유사성을 최대화하기 위해 대조 손실(contrastive loss)을 계산한다. i번째 제안 영역에 대한 대조 손실은 아래 수식 (3)과 같다.

$$L_i = -2 \cdot \frac{<v_i, v'_i>}{\|v_i\|_2 \cdot \|v'_i\|_2} - 2 \cdot \frac{<v_i, v''_i>}{\|v_i\|_2 \cdot \|v''_i\|_2} \quad (3)$$

입력 이미지의 모든 제안 영역에 대한 손실 L은 아래 수식 (4)와 같이 정의된다.

$$L = \frac{1}{K}\sum_{i=1}^{K} L_i \quad (4)$$

여기서 K는 각 이미지에 대한 제안 영역의 개수를 의미한다. 학습 과정에서 Target Network는 직접 가중치를 업데이트 하지 않으며, Online Network로부터 exponential moving average (EMA)로 가중치를 업데이트 받는다. EMA를 통한 가중치 업데이트는 다음 수식 (5)와 같이 나타낼 수 있다.

$$\theta_{target} \leftarrow \tau\theta_{target} + (1-\tau)\,\theta_{online} \quad (5)$$

수식 (5)에서, θ_{target}은 Target Network의 가중치, θ_{online}은 Online Network의 가중치를 의미한다. τ 는 EMA계수이다. SoCo는 이러한 학습 과정을 통해 객체 탐지 작업에서 중요한 객체 수준 표현을 학습하며, 이는 위치와 크기 변화에 강인한 특성을 제공한다.

2. 3장의 실험(표 1) 설정 및 구현 방안

본 연구는 KAIST 데이터셋을 활용하여 실험을 진행하였다. 이 데이터셋은 차량에서 연속적으로 촬영된 주/야간 주행 장면으로 구성되며, 총 95,328개의 컬러 영상(640×480)과 열화상 영상(320×256) 쌍 및 이에 대한 어노테이션으로 이루어져 있다. 데이터셋을 구성하는 주행 영상은 모두 가을(한국의 10월)에 취득되었으며, 도로, 캠퍼스, 도심 환경의 주행 장면으로 구성되었다. 실험은 멀티스펙트럴 데이터를 활용하여 기존 자기 지도 학습 방법론이 해당 데이터에 효과적으로 적용 가능한지를 평가하기 위해 설계되었다.

본 실험에서는 대조 학습 기반 자기 지도 학습 방법론 중 전역적 표현 학습을 중점으로 하는 MoCo [14], BYOL [16], SimSiam [17]과 국소적 표현 학습에 초점을 맞춘 SoCo [21], DetCo [25], CCrop [26]을 비교 대상으로 선정하였다. 실험은 동일한 장면에서 얻어진 컬러 영상과 열화상 영상 쌍을 동일한 임베딩 공간에 투영하는 방식으로 설계되었으며, 이를 통해 각 방법론이 멀티스펙트럴 데이터를 활용하여 학습한 표현이 얼마나 효과적인지를 평가하였다. 사전 학습된 가중치는 검출기의 인코더 네트워크인 ResNet50 [28]을 초기화하는 데 사용되었고, 이후 멀티스펙트럴 보행자 검출기인 MLPD [10]를 활용하여 미세 조정을 진행하였다.

사전 학습 단계에서는 각 방법론의 기본 설정을 유지하며 배치 사이즈(batch size)와 학습 epoch를 조정하였다. MoCo와 SimSiam은 배치 사이즈 256으로 설정하였으며, BYOL은 배치 사이즈 32로 사전 학습을 진행하였다. 국소적 표현 학습 방법론인 DetCo와 CCrop은 각각 배치 사이즈 128과 256으로 설정되었으며, SoCo는 배치 사이즈 42로 설정하여 사전 학습을 진행하였다. 사전 학습 단계에서 모든 방법론은 100epoch 만큼 학습되었다. 각 방법론에 사용된 하이퍼파라미터는 기존 연구에서 제시한 설정을 따랐다. 검출기의 미세 조정 단계에서는 배치 사이즈를 4로 설정하고, 40 epoch 동안 학습을 수행하였다.

표 3의 실험을 제외한 모든 실험(표 1, 2, 4, 5, 6)은 NVIDIA GeForce RTX 2080 Ti 환경에서 진행되었으며, 표 3의 실험은 NVIDIA TESLA V100 환경에서 진행되었다. 성능 평가는 보행자 검출에서 일반적으로 사용되는 Miss Rate를 기준으로 이루어졌다.

3. 4장의 실험(표 2, 3, 4, 5, 6) 설정 및 구현 방안

모든 실험은 SoCo 프레임워크를 기반으로 KAIST 데이터셋을 활용해 수행되었으며, 멀티스펙트럴 검출기는 MLPD가 사용되었다. 표 2, 4, 5, 6의 실험에서는 ResNet50이 MLPD의 인코더 네트워크로 사용되었고, 사전 학습 단계에는 배치 사이즈 42로 400 epoch 학습이 진행되었으며, 미세 조정 단계에는 배치 사이즈 4로 40 epoch 학습되었다. 표 3의 실험은 MLPD의 인코더를 ResNeXt50(32×4d), ResNet50(×2), ResNet101로 교체하여 수행되었다. 각 인코더는 사전 학습 단계에서 배치 사이즈 64로 100 epoch 학습되었고, 검출기의 미세 조정 단계에는 배치 사이즈 4로 40 epoch 학습되었다.

표 2의 실험에서 θ_C는 컬러 영상 도메인만을 활용해 사전 학습하여 얻은 가중치이고, θ_T는 열화상 영상 도메인만을 활용해 사전 학습하여 얻은 가중치이다. θ_{CT}는 컬러 및 열화상 도메인 데이터를 동일 공간에 임베딩하여 대조 학습을 진행해 사전 학습한 가중치이다. '단일 가중치' 초기화 방법은 검출기의 두 입력 인코더를 동일한 가중치로 초기화한 뒤 미세 조정하였고, '모달리티 고려'에서는 검출기의 두 인코더를 θ_C와 θ_T로 각각 초기화한 뒤 미세 조정하였다.

표 3의 실험에서는 표 2의 실험 결과에 대한 경향성을 추가적으로 검증하기 위해, 표 2와 마찬가지로 검출기의 두 입력 인코더를 θ_{CT}로 초기화한 방법과 검출기의 두 인코더를 θ_C 와 θ_T 로 각각 초기화한 방법을 ResNeXt50(32×4d), ResNet50(×2), ResNet101 인코더에서 평가하였다. 백본 교체에 따른 검출 모델의 크기 증가로 표 3의 실험은 NVIDIA TESLA V100 환경에서 분산 학습을 활용해 진행되었으며, 이 외의 검출기 학습 하이퍼파라미터는 MLPD [10]와 동일하다.

표 4의 영역 제안 효과 검증을 위한 패치 생성은 Selective Search 결과를 참고하여 설계되었다. 컬러 영상 데이터에서는 평균 35.68개의 영역이 제안되었기에 107x85 크기의 패치를 36개(6x6) 생성하였으며, 열화상 영상 데이터에서는 평균 13.22개의 영역이 생성되었기에 160x170 크기의 패치를 12개(4x3) 생성하여 학습에 사용하였다. 이 때 범위를 넘어가는 패치 영역은 버림하였다. 랜덤 박스 생성 시에는 박스의 좌표와 크기를 영상의 범위 내부에서 uniform distribution으로 추출해 사용하였다. 이 때 제안 영역의 개수는 컬러 영상 데이터에서 평균 25.46개, 열화상 영상 데이터에서 평균 25.44개 생성해 사전 학습에 활용하였다. 각 영역 제안 방식으로 SoCo 프레임워크를 활용해 사전 학습을 수행하고, 이후 멀티스펙트럴 검출기의 인코더 E_T, E_C를 각각 모달리티 별로 사전 학습된 가중치 θ_T, θ_C로 초기화하여 미세 조정한 검출 결과를 비교하였다.

표 5의 실험에서는 SoCo 프레임워크에서 FPN과 R-CNN Head 없이 특징 맵을 Prediction Layer로 바로 전달하는 방식으로 사전 학습한 가중치와, 본래 SoCo 프레임워크로 사전 학습한 프레임워크를 비교하였다. 두 방식 모두 각 모달리티 별로 사전 학습된 가중치 θ_T, θ_C로 두 입력 인코더 E_T, E_C를 별도로 초기화한 뒤 미세 조정한 결과를 비교하였다.

표 6의 실험에서는 SoCo 사전 학습에 활용되는 제안 영역 및 활용 데이터를 다르게 하였다. P_C는 컬러 영상을 활용해 생성한 제안 영역을, P_T는 열화상 영상을 활용해 생성한 제안 영역을 의미한다. $P_T + P_C$의 경우 P_T와 P_C의 제안 영역을 결합하여 SoCo 사전 학습에 활용하였다. 각 사전 학습 가중치를 활용해 검출기의 인코더 E_T, E_C를 초기화한 후 미세 조정하여 검출 결과를 비교하였다.

VI. 결론

본 논문에서는 기존 자기 지도 학습 연구를 멀티스펙트럴 보행자 검출에 확장 적용하였으며, 실험을 통해 기존 자기

지도 학습 방법에서 멀티스펙트럴 데이터 학습에 긍정적인 영향을 미치는 요소를 분석하였다. 실험 결과, 영상 전역 단위 학습법 보다는 영상의 국소적 특징 학습을 고려한 방법론이 좋은 특징을 학습할 수 있음과, 열화상 영상을 고려해 기존 방법론들을 개선할 필요가 있음을 확인하였다. 또한 각 영상 모달리티를 별도로 사전 학습하여 검출기의 두 인코더를 입력 데이터 특성에 알맞게 별도 초기화하는 것이 검출 성능에 긍정적임을 확인하였다. 특히 기존 자기 지도 학습 방법론 중에서 가장 효과적인 학습 성능을 보인 SoCo 프레임워크를 활용한 추가적인 실험을 통해, 사전 학습 단계에서 Selective Search를 활용한 학습 영역 개선과 다양한 객체 수준 표현 정보를 활용한 학습이 긍정적인 영향을 미침을 확인하였다. 마지막으로 사전 학습 시 활용 모달리티 구성에 대한 실험을 통해 단일 모달리티 영상만을 활용하는 것보다 멀티 모달리티를 모두 활용했을 때 개선된 결과를 보여 멀티스펙트럴 데이터에 대한 자기 지도 학습 적용 이점을 확인하였다.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* vol. 39, no. 6, pp. 1137-1149, 2017.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," *European conference on computer vision (ECCV),* 2016.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-end object detection with transformers," *European Conference on Computer Vision (ECCV),* 2020.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI),* vol. 34, no. 4, pp. 743-761, 2011.

[5] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017.

[6] J. Lee and J. Kang "3D multi-object tracking using instance segmentation of stereo camera images in autonomous vehicles," *Journal of Institute of Control, Robotics and Systems (in Korean),* vol. 30, pp. 1389-1397, Dec. 2024.

[7] Z. Zhang, L. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20-29, 2019.

[8] Y. Zhuang, Z. Pu, J. Hu, and Y. Wang, "Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection," *IEEE Transactions on Network Science and Engineering (TPAMI)*, vol. 9, no. 3, pp. 1282-1295, 2021.

[9] S. M. Hwang, J. S. Park, N. I. Kim, Y. K. Choi, and I. S. Kweon, I. (2015). "Multispectral pedestrian detection: Benchmark dataset and baseline," *IEEE conference on Computer Vision and Pattern Recognition (CVPR),* 2015.

[10] J. W. Kim, H. J. Kim, T. J. Kim, N. I. Kim, and Y. K. Choi, "MLPD: Multi-label pedestrian detector in multispectral domain," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7846-7853, 2021.

[11] J. Liu, S. Zhang, S. Wang, and D. Mataxas, "Multispectral deep neural networks for pedestrian detection," *British Machine Vision Conference (BMVC),* 2016.

[12] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *International Conference on Learning Representations (ICLR)*, 2018.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *International Conference on Machine Learning (ICML)*, 2020.

[16] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, C. Couprie, B. Batteau, L. Blanc, and M. Valko, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] X. Chen, and K. He, "Exploring simple siamese representation learning," *IEEE conference on Computer Vision and Pattern Recognition (CVPR),* 2021.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] U. C. Shin, K. H. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for RGB-thermal semantic segmentation," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[20] J. Bayrooti, N. Goodman, and A. Tamkin, "Multispectral contrastive learning with viewmaker networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[21] F. Wei, Y. Gao, Z. Wu, H. Hu, and S. Lin, "Aligning pretraining for detection via object-level contrastive learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[22] S. J. Ma, and Y. G. Cho, "Robust place recognition using fusion of thermal infrared and RGB cameras," *Journal of Institute of Control, Robotics and Systems (in Korean),* vol. 30, pp. 1414-1421, Dec. 2024.

[23] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[26] X. Peng, K. Wang, Z. Zhu, M. Wang, and Y. You, "Crafting better contrastive views for siamese representation learning," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] X. Glorot, and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 249-256, March 2010.
[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017
[30] S. Zagoruyko, and N. Komodakis “Wide residual networks,” *British Machine Vision Conference (BMVC),* 2016.
[31] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision (IJCV)*, vol. 104, pp. 154-171, 2013.

허 재 연

2024년 세종대학교 지능기전공학부 스마트기기공학전공 졸업. 2024년~현재 세종대학교 대학원 AI로봇학과 석사과정 재학 중. 관심분야는 컴퓨터비전, 인공지능 및 로보틱스.

정 의 철

2018년~현재 세종대학교 지능기전공학부 무인이동체공학과 학사 재학. 관심분야는 컴퓨터비전, 인공지능 및 물체인식.

황 유 진

2022년 세종대학교 지능기전공학부 무인이동체공학전공 졸업. 2024년 세종대학교 대학원 지능기전공학 석사, 2024년~현재 세종대학교 대학원 AI로봇학과 박사과정 재학 중. 관심분야는 컴퓨터비전, 인공지능, 머신러닝.

최 유 경

2006년 숭실대학교 정보통신전자공학과 졸업. 2008년 연세대학교 전기전자공학부 석사, 2018년 한국과학기술원 전기전자공학과 / 로봇학제전공 박사, 2018년~현재 세종대학교 AI로봇학과 교수. 관심분야는 컴퓨터비전, 머신러닝, 로보틱스.