특집논문 (Special Paper)
방송공학회논문지 제28권 제4호, 2023년 7월 (JBE Vol.28, No.4, July 2023)
https://doi.org/10.5909/JBE.2023.28.4.363
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

희소 입력 환경에서의 지식 증류를 활용한 카메라-라이다 센서 융합

황 혜 린[a], 조 동 현[b)‡]

Camera-LiDAR Sensor Fusion using Knowledge Distillation under Sparse Input Environment

Hyerin Hwang[a] and Donghyeon Cho[b)‡]

요 약

최근 자율 주행 기술이 다양한 분야에 접목되면서 자율주행 차량에 관한 기술 발전이 많은 관심을 받고 있다. 자율 주행 차량의 객체 검출 기술은 자율주행 차량의 안전성과 성능 개선에 관한 기술로써 많은 연구가 이루어지고 있다. 주요 인식 센서인 카메라, 레이더, 라이다 중 카메라를 이용한 단일 센서 연구가 일반적이지만 카메라만 이용한 멀티 카메라 기술은 성능 개선에 한계점이 존재한다. 단일 센서의 한계점을 극복하기 위한 방법으로 멀티 모달 정보를 이용한 센서 융합 기술들이 연구되고 있지만, 입력 정보에 따른 성능 차이가 존재한다. 따라서, 본 논문에서는 희소 입력 정보가 들어와도 성능 개선이 가능한 센서 융합을 위한 지식 증류 기법을 제안한다. 제안하는 방법은 교사 모델의 지식을 상대적으로 희소 입력으로 받는 학생 모델로 전달한다. 최종 실험 결과로 희소 입력 정보에도 강인하게 동작하는 카메라-라이다 센서 융합 기반 3D 객체 검출 성능을 검증한다.

Abstract

Recently, as autonomous driving technology has been applied to various fields, technology development related to autonomous vehicles has received a lot of attention. The object detection technology of autonomous vehicles is a technology for improving the safety and performance of autonomous vehicles, and many studies have been conducted. Among the major recognition sensors, camera, radar, and LiDAR, single sensor research using cameras is common, but multi-camera technology using only cameras has limitations in improving performance. Sensor fusion technologies using multi modal information are being studied as a way to overcome the limitations of a single sensor, but there is a performance difference according to input information. Therefore, this paper proposes a knowledge distillation technique for sensor fusion that can improve performance even when sparse input information is received. The proposed method transfers the knowledge of the teacher model to the student model received with relatively sparse input. As a result of the final experiment, we verify the 3D object detection performance based on camera-LiDAR sensor fusion, which operates strongly even with sparse input information.

Keyword : Sensor Fusion, Knowledge Distillation, Object Detection, Deep Learning

Ⅰ. 서 론

자율 주행의 기술 발전으로 많은 산업 분야가 발전 중이다. 대표적으로 자동차 산업은 최근 몇 년 동안 인공지능, 센서 기술 및 컴퓨터 비전의 급격한 발전 덕분에 크게 성장했다. 자율 주행은 차량이 운전자의 개입 없이 스스로 주행 환경을 인식하고 판단하며 제어하는 기술이다. 자율 주행 기술은 사람들의 이동 편의성을 향상하고 교통사고를 줄이며, 교통체증을 완화하는 등 많은 이점을 제공하기 때문에 자율 주행에 관한 연구가 활발하게 진행되고 있다.

자율주행 차량의 초기 연구는 주로 카메라를 이용한 단일 센서 기반의 접근 방식에 집중했다. 그러나 이러한 방식은 다양한 주행 환경과 날씨 조건에 대응하기 어렵다는 한계가 있었다. 이후 연구는 센서 융합(sensor fusion) 기술을 도입하여 카메라, 라이다, 레이더 등의 융합으로 주행 환경에 대한 정보를 정확하게 추론할 수 있도록 연구가 진행 중이다. 예를 들어, 카메라는 풍부한 정보를 포착하고, 라이다는 정확한 공간 정보를 제공하며, 레이더는 즉각적인 속도를 추정한다. 멀티 모달의 센서 융합은 정확한 인식 정보를 위해 매우 중요하다. 따라서 상호보완적인 정보를 통한 센서 융합은 모델의 전체적인 성능을 향상시킨다.

센서 융합의 종류는 크게 세 가지로 나뉜다. 카메라-라이다, 카메라-레이더, 카메라-라이다-레이더 순으로 많이 사용되는 순서이다. 먼저, 카메라-라이다 센서 융합은 카메라에서 높은 해상도와 색상 정보를 얻고, 라이다에서 거리 정보를 받아 객체 검출의 성능을 올릴 수 있으며, 라이다의 거리 정보를 활용하여 카메라에서 발생하는 객체에 대한 거리 측정의 어려움을 보완한다. 카메라-레이더의 센서 융합은 레이더 센서의 낮과 밤 및 다양한 날씨 조건에서도 성능이 안정적이어서 카메라 환경에 따른 성능 저하를 보완한다. 마지막으로 카메라-라이다-레이더의 센서 융합은 세 가지 센서의 정보를 종합하여 객체 인식, 추적, 및 위치 추정의 정확도를 향상시킬 수 있다. 하지만, 모델의 복잡성이 증가하여 실용적으로 적용하기 어렵다는 단점이 있어 연구가 많이 필요한 분야이다.

카메라-라이다 센서 융합은 카메라의 높은 해상도와 색상 정보, 라이다의 높은 정확도와 거리 정보를 사용하여 객체 검출의 안정성을 높인다. 기존 카메라-라이다 센서 융합 방법은 result-level, proposal-level, point-level의 세 가지에서, 최근 BEV(Bird's-eye-view) fusion의 방식이 추가되었다. Result-level method는 3D proposal을 얻기 위해서 2D detector를 사용하는 방식이다. 그리고 proposal-level fusion method는 다중 센서에서 생성된 region proposal에 RoIPool을 적용해 결합하는 방식이다. 최신 연구들의 카메라-라이다 센서 융합 방법은 point-level fusion, BEV fusion의 두 가지 방법을 꼽을 수 있다. 그림 1에서 보듯, point-level 센서 융합 방법은 calibration matrix를 기반으로 라이다 point와 image 픽셀 간의 강한 연관성을 찾은 다음, point 간의 연결을 통해 관련 픽셀의 segmentation score를 구하거나, CNN feature로 라이다 feature를 augmentation 하는 방법이 있다. 하지만, 실제 주행 환경에선 calibration이 부

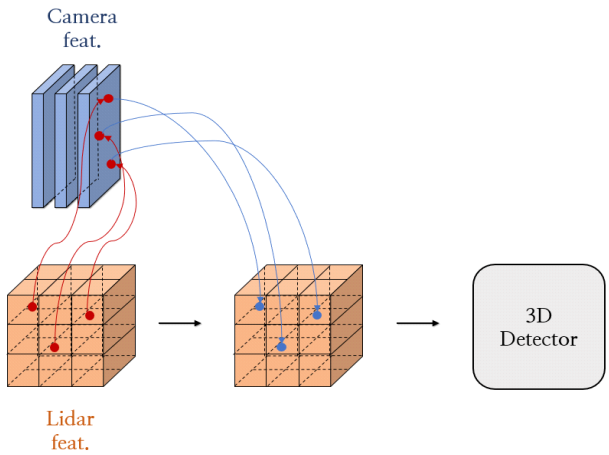

그림 1. Point-level fusion 방식의 카메라-라이다 센서 융합 방법
Fig. 1. Camera-LiDAR sensor fusion method of point-level fusion

a) 충남대학교 컴퓨터공학과(Department of Computer Engineering, Chungnam National University)
b) 충남대학교 전자공학과(Department of Electronics Engineering, Chungnam National University)
‡ Corresponding Author : 조동현(Donghyeon Cho)
E-mail: : cdh12242@cnu.ac.kr
Tel: +82-42-821-5667
ORCID: ttps://orcid.org/0000-0002-2184-921X
※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A4A1032580).
※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021 R1A4A1032580).
· Manuscript May 30, 2023; Revised July 6, 2023; Accepted July 6, 2023.

정확하거나 다시 수행해야 하는 경우가 있기에 취약점이 존재한다. 따라서, 본 연구는 BEV fusion의 센서 융합 방식을 채택한다. BEV fusion 방법은 각 센서에서 BEV feature로 변환하여 BEV feature를 결합해주는 방식이다. 구체적으로 그림 2와 같이 카메라 센서와 라이다 센서의 raw 입력을 같은 BEV feature 공간으로 인코딩한다. 그런 다음 카메라의 인코딩된 입력과 라이다의 인코딩된 입력이 융합하고 3D detection head로 전달되어 객체 검출을 수행한다.

지식 증류(knowledge distillation)란, 큰 모델의 지식을 작은 모델로 전달하는 기술이다. 이는 일반적으로 서로 다른 선생 모델(teacher model)과 학생 모델(student model)을 사용하여 이루어지며, 큰 모델이 학습한 특정 지식을 작은 모델로 전달하여 작은 모델의 성능을 향상 시킨다. 대표적인 주요 기술로는 선생 모델의 출력을 사용하여 학생 모델을 학습하는 teacher-student 학습과, 모델이 자기 자신으로부터 지식을 추출하여 전달하는 self-distillation이 있다. Self-distillation은 일반적인 teacher-student의 지식 증류와 달리, 추가적인 데이터나 다른 선생 모델이 필요하지 않기 때문에 실행 비용의 절감을 장점으로 꼽을 수 있다.

본 연구는 다중 모달리티 모델이 다른 센서로부터 지식을 학습할 수 있지만 높은 비용이 발생하는 문제점을 해결하고자 self-distillation 기법을 도입한다. 이는 희소한 데이터가 들어올 때 강인한 객체 검출을 가능하게 한다. 이전의 연구들은 단일 모달리티에 지식 증류를 적용하여 네트워크가 간단하고 적은 비용으로 구현했다. 그러나 같은 모달리티 간의 지식증류를 적용하는 것은 제한된 모달리티 사용으로 다양한 환경에서의 활용이 제한될 수 있으므로 성능 개선에 한계가 존재한다. 또한, 다중 모달리티에 대한 teacher-student 학습을 활용하는 것은 각기 다른 모달리티 간의 학습을 가능하게 하는 이점을 가지고 있지만, 데이터 처리에 필요한 비용과 병렬 처리로 인한 오버헤드, 지식 변환의 복잡성으로 인해 최적화가 어려워진다는 단점이 있다. 그리고 teacher-student 학습은 학습 모델에 따라 훈련 성능의 차이가 크게 발생한다.

이에 대한 해결책으로, 본 연구에서 기여하는 바는 다음과 같다.

(1) 학생 모델과 선생 모델을 같은 구조로 사용하는 self-distillation을 이용하여, 다중 모달리티 모델에서도 낮은 비용으로 사용 가능하다.
(2) 입력 데이터가 희소하더라도 BEV(Bird's-eye-view) feature map에 지식 증류를 적용하면 강인한 결괏값을 얻을 수 있다.
(3) 같은 모델로 지식 증류를 적용한 점을 활용하여 사전 학습된 하나의 선생 모델 결괏값에 여러가지 설정의 학생 모델을 학습시킬 수 있다.

BEV(Bird's- eye-view) feature map으로 융합하는 구조에 각 센서 별로 희소한 데이터가 입력으로 들어왔을 때

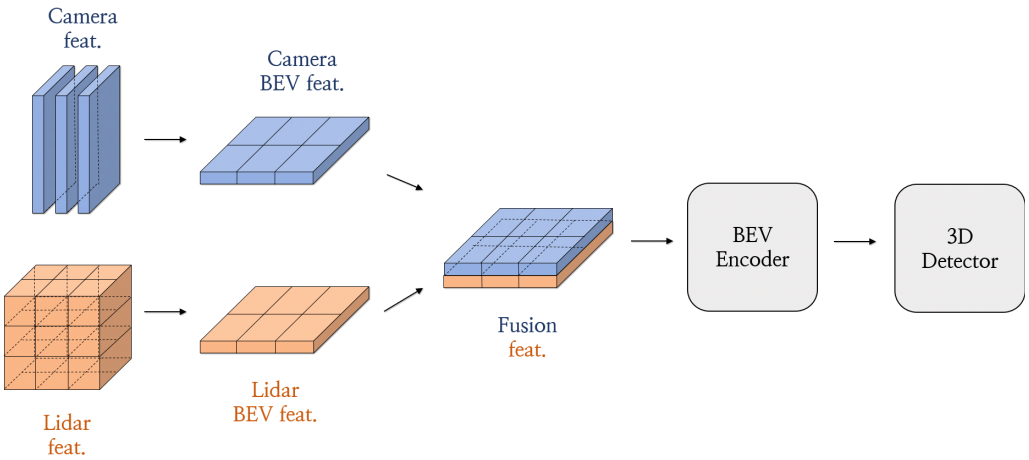

그림 2. BEV fusion 방식의 카메라-라이다 센서 융합 방법
Fig. 2. Camera-LiDAR sensor fusion method of BEV fusion

self-distillation을 적용하는 방법은 현존하지 않는다. 본 연구에서는 희소한 입력 데이터가 들어왔을 때 지식 증류를 사용하여 개선하는 방법을 제안한다. 실제 주행 환경에서는 입력 데이터가 불안정하기 때문에 이러한 환경을 가정하게 되었다. 각 센서를 희소하게 만들어 성능이 좋지 않을 때, 지식 증류의 적용으로 강인하게 동작하는 것을 보여준다. 따라서, 성능을 올리기 위해 카메라-라이다 센서 융합 구조에 지식 증류를 적용한 방법은 그림 3과 같이 나타낸다.

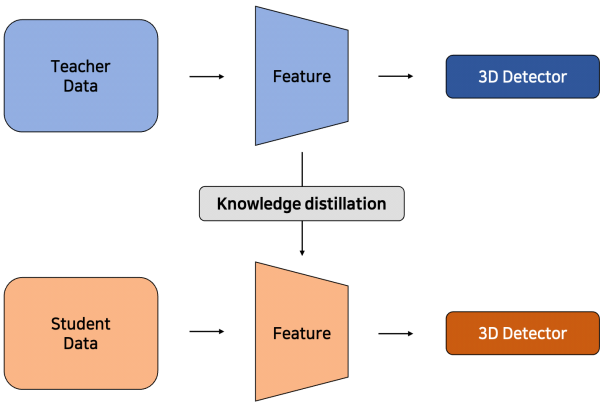

그림 3. 지식 증류 방법
Fig. 3. Knowledge distillation method

본 논문은 다음과 같이 구성한다. Ⅱ절에서 관련 연구를 통한 최신 연구 동향 및 개념에 대해 설명한다. Ⅲ절과 Ⅳ절에서는 제안하고자 하는 방법과 그에 따라 성능 향상이 되었음을 제시한다. Ⅵ절에서는 연구의 주요 결과 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

1. Camera-LiDAR Sensor Fusion

일반적으로 카메라와 라이다에서 feature들이 상호 보완되는 정보를 갖고 있기 때문에 두 가지의 센서를 융합하는 연구가 개발하기 시작했다. 이는 3D 객체 검출에서 표준 방법이 되었으며, 최신 연구로 나오는 카메라-라이다 기반의 detector들은 라이다 point와 image의 정보를 융합하여 단일 모달리티 기반의 detector보다 우수한 성능을 보인다. PointPillars[1]는 3D 객체 검출을 위해 카메라 RGB 입력을 2D 픽셀 그리드로 변환하고, 라이다 데이터를 3D point 피라미드로 변환한 후, 두 데이터를 함께 처리하여 객체를 검출한다. 카메라-라이다 센서 융합의 방법인 result-level Fusion을 적용한 PointNet[2]은 2D 검출 결과를 통해 라이다 point cloud와 해당 point에 대응하는 image 정보를 함께 활용하여 객체를 검출하는 방법이다. Proposal-level Fusion은 proposal 영역을 생성한 뒤 각 proposal 영역은 객체가 있을 가능성이 있는 위치를 나타낸다. 이후 proposal 영역들은 후속 처리 단계에서 분류(classification)나 위치 조정(regression)을 통해 실제 객체의 class와 위치를 예측하는 방법이다. 반면에 point-level fusion 방법은 image feature를 point로 변환하여 라이다의 point와 fusion하는 방법이다. Point-level fusion이 처음 제안된 PointPainting[3]은 라이다를 image에 projection하고, segmentation 정보를 라이다 정보에 concatenation하고 segmentation 정보를 붙인 라이다 point cloud를 기존의 라이다 기반 아키텍처에 넣어준다. 이후 나온 PointAugmenting[4] 은 사전 훈련된 2D 객체 감지 모델에서 추출된 point 별 CNN feature를 사용하여 point cloud를 decoration하고, 이를 기반으로 3D 객체 감지를 수행하면서 기존 방법보다 높은 성능을 보였다.

최신 연구인 FUTR3D[5]는 query 기반 모달리티에 트랜스포머 디코더를 사용하여 end-to-end의 3D 객체 검출을 수행한다. TransFusion[6]은 3D 공간에서 object query를 정의하고 이러한 proposal에 image feature를 융합한다. MV3D[3]은 object proposal을 3D로 생성한 다음 센서 융합을 수행한다. DeepInteraction[8]은 고유한 특성을 활용할 수 있도록 각 모달리티의 representation을 학습하고 유지하는 새로운 interaction 전략을 사용한다. BEV-level Fusion 방법이 나오면서 최근 연구 중 가장 높은 성능을 보여준다. BEV-level Fusion 방법은 BEV feature map을 융합하는 방식이다. 먼저 제안된 BEVFusion[9]은 camera stream이 라이다 데이터의 입력 데이터에 의존하지 않는 방식을 제안했다. 이후 나온 연구인 BEVFusion[10]은 image feature와 라이다 feature를 모두 BEV로 변환하여 두 모달리티 간 융합 및 결과 예측을 제안한다. UVTR[11]은 transformer를 사용하여 3D 객체 검출을 위해 복셀 공간에서 다중 모달리티의 표현(representation)을 통합한다.

2. Knowledge Distillation

지식 증류(knowledge distillation)는 모델 압축을 위해 [12]에서 처음 제안되었다. 해당 논문은 선생 모델의 네트워크에서 학습된 지식을 학생 모델에게 전달하는 방법을 소개했다. 이후 다양한 연구에서 지식 증류 기법을 개선하여 응용되었다. 지식 증류는 컴퓨터 비전의 2D 객체 탐지 및 segmentation 분야 등으로 다양하게 발전해나갔다. 최근에는 detector로 지식을 전달하기 위해 3D 객체 탐지에 도입되었다. 단일 모달리티 detector로 지식 증류를 전달하는 방법인 Monodistill[13]는 라이다를 활용하여 3D 공간을 projection한 카메라 기반 선생 모델을 훈련시켜, 카메라 기반 학생 모델 detector에 라이다 point의 깊이 정보를 전달하는 것을 제안한다. 최신 다중 모달리티를 사용하기 위해 제안된 UVTR[11]은 복셀 공간에서 다중 모달리티의 표현(representation)을 통합될 때, 복셀 인코더 과정에서 지식 증류를 적용한다. 하지만 이 방법들은 선생 모델과 학생 모델의 모달리티가 제한된다. 대신하여 나온 UniDistill[14]는 선생 모델과 학생 모델의 모달리티를 고정하지 않고 4개의 경로를 제안하여, 개선된 성능을 보여준다.

본 논문은 현재 최고 성능이면서 다중 모달리티를 사용한 BEVFusion[10]을 기반으로 채택했다. 포괄적인 feature map에 지식 증류를 적용하는 UniDistill[14]와는 달리, 3가지의 BEV feature map 부분에 각각 지식 증류를 적용하여 센서별 성능 향상을 비교한다. 또한, self-distillation을 이용하여 합리적인 실행 비용으로 효율적인 결과를 제공한다.

Ⅲ. 제안 방법

이 장에서는 센서 융합 모델에 지식 증류를 적용한 방법을 설명한다.

1. Problem Setting

본 연구에서 선생 모델(teacher model)과 학생 모델(student model)은 동일한 모델을 사용했다. 선생 모델에서 학습한 결과를 학생 모델에게 지식을 전달하여 학생 모델이 선생 모델의 출력값에 가까워지도록 사용한다. 선생 모델과 학생 모델에서 사용한 모델은 BEV feature map이 융합하는 구조다. 카메라의 BEV feature map, 라이다의 BEV feature map, 카메라-라이다가 융합된 BEV feature map에 각각 지식 증류를 적용하여 성능을 비교한다. 연구에서 BEVFusion[10]의 구조를 기반으로 연구를 진행하였다. 선생 모델은 BEVFusion[10]의 최고 성능이 나온 기존 옵션을 그대로 사용하여 사전 학습된 모델을 선생 모델로 채택하였다. 사전 학습된 하나의 선생 모델 결괏값에 여러 버전의 학생 모델을 검증하였다. 학생 모델은 선생 모델과 같은 구조를 가지지만, 희소하거나 품질이 낮은 입력을 사용한다. 학생 모델이 학습될 때 사전 학습된 선생 모델의 지식을 전달받는다. 본 논문에서 다루는 problem setting은 그림 4와 같다.

제안하는 네트워크는 센서 융합의 구조에서 지식 증류를 적용했을 때 객체 검출의 성능이 올라가도록 PyTorch 라이브러리로 구성했다. 기존 네트워크는 카메라로 RGB image와, 라이다에서 point 입력을 받아 각각의 처리 단계를 거친 후 입력들이 융합되어 객체 검출한다. 해당 네트워크는 학생 모델에서 희소 입력 데이터로 들어왔을 때 중간 단계에 지식 증류를 적용하여 객체 검출의 성능을 높이고자 고안되었다. Camera BEV feature, LiDAR BEV feature, fusion BEV feature인 feature map에 한 부분씩 지식 증류를 적용하여 어떤 feature map에 지식 증류를 적용했을 때 성능 개선이 높은지 비교한다.

그림 4에서 camera BEV feature, LiDAR BEV feature, fusion BEV feature에 지식 증류를 적용한 것을 나타낸다. 입력 센서는 카메라의 원본 데이터 혹은 희소한 데이터, 라이다의 원본 데이터 또는 희소한 데이터 중 한 가지를 원본으로 사용하고, 다른 한 가지 센서는 희소하게 데이터를 활용하도록 설정한다. 예를 들어, 카메라의 입력 데이터가 희소하게 들어갔을 때, 카메라의 feature map에 지식 증류를 적용하였다. 이때, 라이다는 원본 데이터가 입력으로 사용되고, 라이다의 feature map는 지식 증류를 사용하지 않는다. 반대로 라이다의 입력 데이터가 희소한 경우에 대해서도 동일한 방식으로 적용된다. 추가로, 라이다의 입력 데이터가 희소하게 들어갔을 때 융합된 feature map에 지식 증

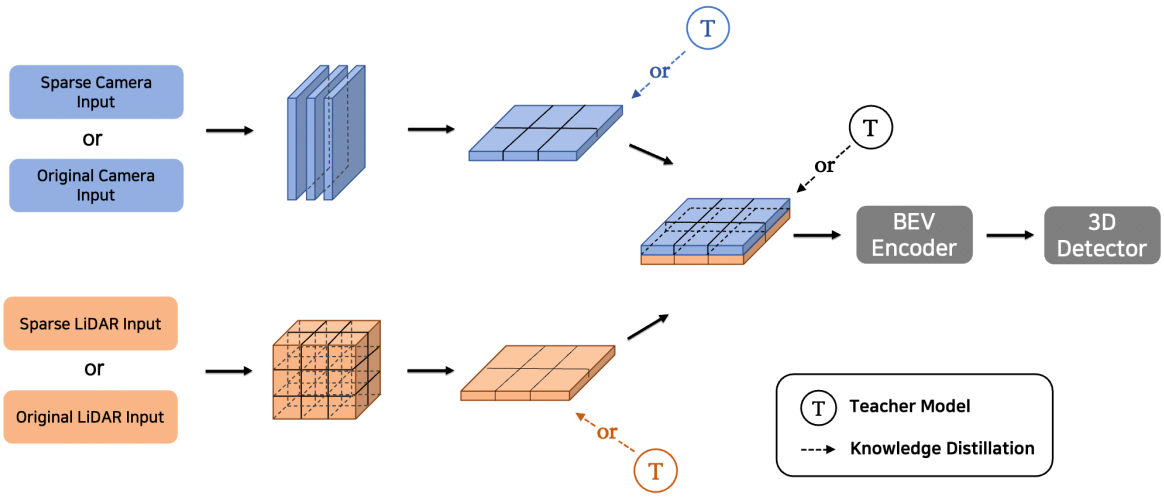

그림 4. 지식 증류를 적용한 카메라-라이다 센서 융합의 구조
Fig. 4. Structure of Camera-LiDAR sensor fusion with knowledge distillation

류를 적용하여 성능을 향상시키는지 비교하였다. 각 실험은 중복 없이 한 가지 feature map에만 지식 증류를 적용한다. 따라서, 학생 모델 구조에서 선생 모델을 불러와 비교하고자 하는 feature map 단계에 지식 증류를 적용하여 어느 feature map을 지식 증류를 적용한 것이 좋은지 비교한다.

2. Data Preparation

본 연구는 희소 입력이 들어왔을 때 3D 객체 검출 성능을 올리는 것을 목표로 한다. 카메라의 데이터 품질을 떨어뜨리기 위해 카메라 입력의 매개변수를 수정하는 방법을 적용했다. 위 코드는 SwinTransformer[15]로 사전 학습된 backbone 파일을 불러와 초기 가중치를 사용하기 때문에 매개변수를 조절하는 방법을 채택했다. 기존 입력 채널수를 256에서 200로 줄이는 방식을 이용해 카메라 정보를 다 사용하지 못하여 품질을 떨어지도록 설정했다. 마찬가지로 학생 모델이 입력으로 사용할 희소 라이다 데이터를 생성하기 위해, 라이다 입력 데이터의 형식인 복셀을 조절하여 실험을 진행하였다. 선생 모델의 복셀 사이즈와 복셀 최대 개수는 각 [0.075, 0.075, 0.2]와 [120,000, 160,000]다. 여기서 복셀 사이즈인 [0.075, 0.075, 0.2]는 3차원 복셀의 X축 방향, Y축 방향, Z축 방향을 나타낸다. 복셀 최대 개수를 나타내는 [120,000, 160,000]은 2차원 공간에서의 복셀 그리드의 가로 방향과 세로 방향의 복셀 개수를 나타낸다. 복셀 사이즈가 클수록, 복셀 최대 개수가 적을수록 희소한 입력이다. 학생 모델에 사용되는 복셀 사이즈와 복셀 최대 개수는 [0.1,

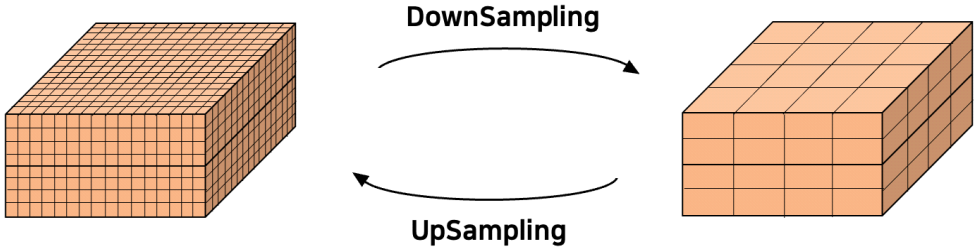

그림 5. 복셀 그리드의 샘플링에 따른 변화
Fig. 5. Variation of voxel grid with sampling

0.1, 0.25], [100,000, 140,000]과 [0.09, 0.09, 0.2]와 [110,000, 150,000]로 설정했다.

복셀 사이즈가 커지고 복셀 최대 개수가 적어질수록 정보가 간소화되고 축소되므로 정보 손실이 발생한다. 그림 5는 복셀 사이즈와 복셀 최대 개수를 줄였을 때 복셀 그리드의 변화를 보여준다. 구체적으로 복셀 사이즈를 키우고 최대 복셀 개수가 적어지면 같은 크기의 그리드 안에서 정보가 축소되고, 복셀의 개수 제한으로 인해 더 많은 공간 정보를 얻지 못하여 정확도가 떨어진다. 따라서, 희소한 입력 데이터를 사용하기 위해 다운 샘플링(downsampling)을 통한 출력을 실험에 적용한다. 희소한 입력 데이터의 환경을 가정하였을 때 성능 개선을 확인하기 위한 방법으로 위와 같은 방법을 사용하였다.

3. Knowledge Distillation Loss

학습 모델을 학습하기 위해 옵티마이저인 AdamW[16]을 사용하였다. 먼저 L2 regularization은 손실함수에 weight에 대한 제곱을 해줌으로써 과적합(overfitting)을 방지한다. 그리고 weight decay는 gradient descent에서 weight 업데이트를 할 때, 이전 weight의 크기를 일정 비율 감소시켜 과적합을 방지한다. AdamW는 앞서 두 가지 방법 중 weight decay을 사용하여 과적합을 방지하는 Adam의 변형이다.

지식 증류 손실 함수값을 계산하기 위한 손실 함수로 MSE loss(mean squared error)와 L1 loss(mean absolute error)를 사용하여 결괏값을 비교하였다. MSE는 예측값과 실제 값 차이의 제곱을 평균한 값이며 주어진 데이터 point에 대한 MSE는 다음과 같이 계산한다.

$$MSE = (1/n) * \Sigma (y_i - \hat{y}_i)^2 \quad (1)$$

여기서, y_i는 실제 값, $\hat{y}_i$는 예측값, n은 데이터 point의 수를 의미한다. MSE를 최소화하는 것은 예측값과 실제 값 간의 전반적인 차이를 줄이는 것을 목표로 사용한다.

L1 Loss는 예측값과 실제 값 차이의 절댓값 평균이다. 주어진 데이터 point에 대한 L1 Loss는 다음과 같이 계산한다.

$$L1 = (1/n) * \Sigma |y_i - \hat{y}_i| \quad (2)$$

L1 Loss는 오차의 절댓값을 사용하므로 오차의 크기에 상관없이 모두 동일한 가중치로 처리한다. L1 Loss를 최소화하는 것은 예측값과 실제 값 간의 전반적인 차이를 줄이는 것을 목표로 사용한다.

MSE Loss를 지식 증류 함수의 기본 loss로 사용했으며, loss에 따른 결과를 비교하기 위해 L1 Loss를 사용했다. 선생 모델의 feature map을 $F_{Teacher}$, 학생 모델의 feature map을 $F_{Student}$로 표기한다. Feature map을 비교하기 위해 지식 증류 손실 함수 계산 방법은 다음과 같다.

$$KDLoss(MSE) = (1/n) * \Sigma (F_{Teacher} - F_{Student})^2 \quad (3)$$

$$KDLoss(L1) = (1/n) * \Sigma |F_{Teacher} - F_{Student}| \quad (4)$$

앞서 구해진 지식 증류 함수는 기존 함수에 더해주는 방식을 적용하여 학생 모델이 선생 모델의 출력값을 따라가도록 구성한다. 최종 손실 함수는 아래와 같다.

$$TotalLoss = DetLoss + KDLoss \quad (5)$$

이때, $DetLoss$는 기존의 객체 검출의 손실 함수이며, $KDLoss$는 지식 증류 손실 함수를 의미한다. 이와 같이 구해진 $TotalLoss$를 통해 학생 모델은 선생 모델과의 차이를 줄이는 방식으로 학습을 진행한다.

Ⅳ. 실험 결과

1. 데이터 세트

우리는 실험평가를 위해 nuScenes 데이터 세트[17]를 사용했다. nuScenes는 자율 주행 차량 관련 연구를 위한 공개 데이터 세트이며, Waymo 데이터 세트[18]와 KITTI와 같은 다른 자율 주행 관련 데이터 세트들과 함께 가장 널리 사용되는 데이터 세트 중 하나이다. 자율 주행 시스템의 개발과 성능 평가를 위해 사용된다. 이 데이터 세트는 6개 카메라의 image, 5개 radar의 point, 1개 라이다의 image와 함께

총 1,000개의 장면이 포함되어 있다. 각 장면은 약 20초 동안의 시간 동안 수집된 데이터를 나타내며 대략 40,000개의 연속적인 프레임을 포함한다. 본 연구에서는 nuScenes 데이터 세트의 센서 중 카메라와 라이다 데이터만 사용하여 연구를 진행하였다.

2. 평가 매트릭

mAP(mean average precision)와 NDS(nuScenes detection score)를 사용한다. mAP는 object detector의 정확도를 측정하여 모델의 성능을 평가하는 지표이다. 주로 IOU를 threshold로 사용하여 location, size, orientation을 구분하지 않고 한 번에 평가가 이루어진다. NDS란 nuScenes 데이터 세트의 객체 탐지 성능을 평가하기 위해 사용되는 지표이다. NDS는 mAP, mATE, mASE, mAOE, mAVE, mAAE의 가중 합을 계산하는 것이다. mAVE, mAOE, mATE는 1보다 클 수 있고, 나머지는 0과 1 사이의 값을 갖는다.

NDS는 식 (6)과 같이 계산한다.

NDS의 값이 높을 수록 모델의 정확도와 완전성이 높다는 것을 나타낸다.

mAP은 식 (7)과 같이 계산한다.

$$mAP=(1/C)*\Sigma(AP_c) \tag{7}$$

여기서 C는 class의 수를 의미하며, AP_c는 각 class c에 대한 평균 Precision을 나타낸다. AP_c을 모든 class에 대해 합산한 후, class 수로 나누어 평균을 계산한다. mAP는 보통 0부터 1까지의 값으로 나타내며, 값이 높을 수록 모델의 탐지 성능이 좋다는 것을 의미한다.

3. 정량적 결과 분석

표 1, 표 2는 knowledge distillation의 적용 위치에 따른 결과를 보여준다. 표 3은 knowledge distillation loss의 비교 결괏값을 보여주며, 표 4는 라이다 데이터 품질 정도에 따른 결괏값이다. 희소 입력 데이터가 들어왔을 때를 가정하여 학습을 진행하였으며, 학생 모델의 기본 고정값을 mAP 55.58과 NDS 60.31로 낮추어 진행하였다. 학생 모델이 mAP 55.58과 NDS 60.31에서 성능 향상이 얼마나 이루어지는지 확인한다. 표 1은 MSE loss를 이용한 각 feature map 위치에 따라 지식 증류를 적용한 결과표이다. 표 1, 표 2의 복셀 사이즈와 복셀 최대 개수는 [0.1, 0.1, 0.25]와 [100,000, 140,000]이며, 표 3에서 사용된 복셀 사이즈와 복셀 최대 개수는 [0.09, 0.09, 0.2]와 [110,000, 150,000]이다. 표 1에서 라이다 feature map, 융합된 feature map이 지식 증류로 인한 성능 향상이 된 것을 보여준다. 표 2에서는 fusion feature map에 대해 지식 증류 손실 함수 종류에 따른

$$NDS=\frac{1}{10}(5mAP+mATE+mASE+mAOE+mAVE+mAAE) \tag{6}$$

표 1. 라이다의 입력 품질을 낮춘 후 지식 증류 적용 성능 비교
Table 1. Comparison of knowledge distillation application performance after lowering the input quality of LiDAR

Method	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
Original	55.58	60.31	34.86	27.52	50.99	41.69	19.71
KDLoss(MSE) + LiDAR feature	56.50	62.07	33.995	26.70	45.85	36.76	18.50
KDLoss(MSE) + Fusion feature	58.21	63.01	32.77	26.83	43.57	39.07	18.77

표 2. Fusion feature map의 지식 증류 Loss 사용에 따른 성능 비교
Table 2. Comparison of performance using knowledge distillation loss in fusion feature map

KD Loss	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
MSE loss	58.21	63.01	32.77	26.83	43.57	39.07	18.77
L1 loss	57.15	62.30	33.39	26.83	47.40	36.09	19.08

표 3. 다른 라이다 데이터 품질에 대한 지식 증류 적용 결과
Table 3. Results of applying knowledge distillation to different LiDAR data quality

Method	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
Original	60.76	63.36	32.10	26.76	52.27	39.90	19.22
KDLoss(MSE) + Fusion feature	60.96	64.90	31.47	26.16	46.42	32.89	18.93
KDLoss(L1) + Fusion feature	61.27	65.40	31.56	26.01	44.91	30.98	18,92

표 4. 카메라의 입력 품질을 낮춘 후 지식 증류 적용 성능 비교
Table 4. Comparison of knowledge distillation application performance after lowering the input quality of Camera

Method	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
Original	62.64	66.24	29.40	26.25	43.88	32.01	19.20
KDLoss(MSE) + Camera feature	62.94	67.23	29.46	26.15	36.77	30.32	19.62
KDLoss(L1) + Camera feature	63.11	67.76	30.08	26.03	33.46	29.10	19.28

결과 차이이며, fusion feature map의 경우에는 L1 loss보다 MSE loss를 사용했을 때 성능이 높은 것을 확인할 수 있다.

표 3에서의 복셀 사이즈와 복셀 최대 개수를 조절했을 때를 비교한다. 라이다 데이터 품질의 정도에 따른 차이를 확인하기 위한 결과표이다. 복셀 사이즈와 복셀 최대 개수를 각 [0.09, 0.09, 0.2], [110,000, 150,000]로 조정하여 지식 증류를 적용했을 때와 하지 않았을 때를 비교한다. 앞선 실험과 같이 batch size, learning rate, epoch 등 하이퍼파라미터를 동일하게 맞춰 진행하였다. 기존 각 표에서 볼 수 있듯이 Knowledge Distillation Loss를 적용하면 학생 모델의 기본 고정값인 mAP 55.58과 NDS 60.31보다 더 높은 mAP 60.76, NDS 63.36으로 진행하여 지식 증류가 적용했을 때도 높은 성능을 얻은 것을 확인할 수 있다.

추가적으로 앞선 표 1, 표 2, 표 3과 달리 표 4에서는 카메라의 입력 품질을 낮춘 후 지식 증류를 적용했을 때의 결과이다. 카메라의 품질을 낮췄을 때 학생 모델의 기본 고정값은 mAP 62.64, NDS 66.24이다. 카메라에서도 희소한 입력 데이터가 들어올 때 지식 증류를 적용하면 성능 개선이 가능하다. 라이다에 희소한 입력 데이터에 지식 증류를 적용했을 때 보다 성능 향상이 낮지만, 카메라 센서도 희소한 입력 데이터 부분에 지식 증류를 적용하면 성능 향상이 있음을 보여준다.

4. 정성적 결과 분석

그림 6은 nuScenes 데이터 세트를 이용하여 fusion fea-

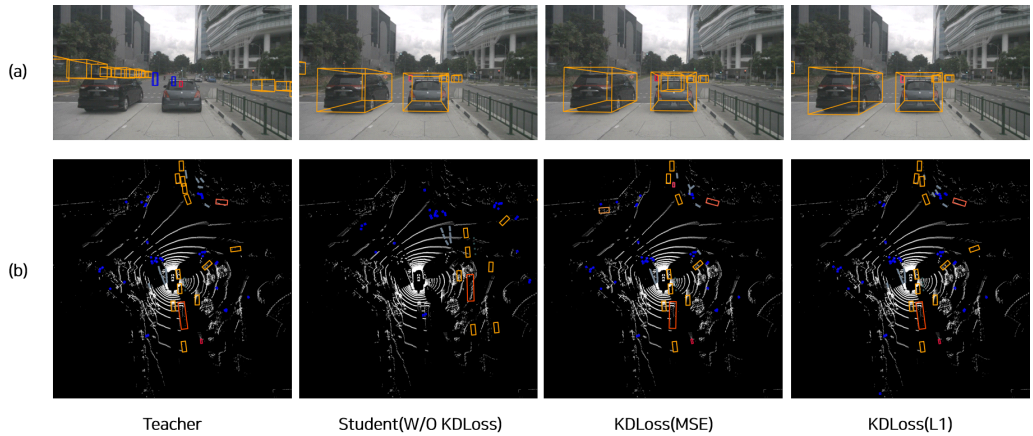

그림 6. Fusion feature map에 대한 정성적 결과
Fig. 6. Qualitative results for fusion feature map

ture map에 지식 증류를 적용한 결과다.

카메라와 라이다의 시각화 결과 이미지에 대하여 선생 모델과 지식 증류를 학생 모델에 적용했을 때를 각각 제시하였다. 먼저, (a)는 카메라에 대한 시각적 결과이다. 선생 모델은 주로 작은 객체에 대한 객체 탐지가 잘 이루어지는 것을 확인할 수 있다. 이로 인해 높은 정확도의 성능을 갖는다. 희소한 데이터가 입력으로 들어간 학생 모델의 경우, 선생 모델과 달리 큰 객체에 관한 객체 탐지가 주로 이루어진다. 학생 모델에 지식 증류 손실함수를 적용한 이미지는 MSE 손실함수를 사용했을 때 객체 탐지 성능이 약간 향상한 것을 알 수 있다. L1 손실함수를 사용했을 때는 학생 모델과 비슷한 성능으로 객체 탐지가 이루어졌다. 이는 지식 증류를 통한 성능 향상이 실제 정성적 결과로는 비슷하거나 약간의 상승만 있는 것으로 보인다. (b)는 라이다에 대한 시각적 결과이다. 자세히 살펴보면 여러 가지 class의 객체 탐지가 이루어진 것을 확인할 수 있다. 먼저, 선생 모델의 경우 카메라에 대한 시각적 결과와 비슷하게 작은 객체에 대해 객체 탐지가 잘 이루어진다. 학생 모델의 경우 전체적으로 객체 탐지가 잘 이루어졌지만 선생 모델에 비해 객체 탐지의 성능이 떨어진다. 학생 모델에 지식 증류를 적용한 결과 이미지는 성능 향상한 것을 보여준다. 결과적으로 지식 증류를 적용한 학생 모델의 시각화 결과를 보면 희소한 입력 데이터에도 불구하고 성능 개선이 이루어지면서 객체 검출이 조금은 향상되었다는 점을 관찰할 수 있다. 학생 모델은 선생 모델의 지식을 전달받아서 훈련되었고, 이는 지식 증류가 학생 모델에게 선생 모델의 지식과 일반화 능력을 전달하며, 효과적으로 성능을 개선시키는 역할을 한다는 것을 시사한다. 다만 성능 향상이 크지 않아 실제 주행 환경에서는 효과적인 성능 개선이 부족하다는 것을 확인할 수 있다.

Ⅴ. 결 론

본 연구는 센서 융합을 이용한 모델 구조에 지식 증류를 적용하여 성능 차이를 비교하였다. 학생 모델에 희소 입력 데이터가 들어왔을 때 선생 모델 지식을 전달해주어 선생 모델의 학습 결과를 따라가면서 객체 검출의 성능을 높였다. 하지만 입력된 데이터 품질이 현저히 떨어질 경우, 미검출 및 오검출이 올라가 성능이 달라지므로 적당한 품질의 데이터가 필요하다는 단점이 있다. 향후 연구에서 카메라-라이다-레이더 센서 융합하여 희소 입력 데이터가 들어왔을 때 지식 증류를 적용하여 성능 개선을 하는 연구를 진행할 예정이다.

참 고 문 헌 (References)

[1] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang and O. Beijbom, "PointPillars: Fast Encoders for Object Detection From Point Clouds," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 12697-12705, 2019.
doi: https://doi.org/10.1109/CVPR.2019.01298

[2] R. Q. Charles, H. Su, M. Kaichun and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 652-660, 2017.
doi: https://doi.org/10.1109/CVPR.2017.16

[3] S. Vora, A. H. Lang, B. Helou and O. Beijbom, "PointPainting: Sequential Fusion for 3D Object Detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 4604-4612, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00466

[4] C. Wang, C. Ma, M. Zhu and X. Yang, "PointAugmenting: Cross-Modal Augmentation for 3D Object Detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 11794-11803, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.01162

[5] Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H, "Futr3d: A unified sensor fusion framework for 3d detection," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 172-181, 2023.
doi: https://doi.org/10.48550/arXiv.2203.10642

[6] Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C. L, "TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 1090-1099, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.00116

[7] X. Chen, H. Ma, J. Wan, B. Li and T. Xia, "Multi-view 3D Object Detection Network for Autonomous Driving," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1907-1915, 2017.
doi: https://doi.org/10.1109/CVPR.2017.691

[8] Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L, "Deepinteraction: 3d object detection via modality interaction," Advances in Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, pp. 1992-2005, 2022.

doi: https://doi.org/10.48550/arXiv.2208.11112

[9] Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, Z, "BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework," Advances in Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, pp. 10421-10434, 2022.
doi: https://doi.org/10.48550/arXiv.2205.13790

[10] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," IEEE International Conference on Robotics and Automation (ICRA), 2023.
doi: https://doi.org/10.48550/arXiv.2205.13542

[11] Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J, "Unifying voxel-based representation with transformer for 3d object detection," Advances in Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, pp. 18442-18455, 2022.
doi: https://doi.org/10.48550/arXiv.2206.00630

[12] Hinton, G., Vinyals, O., Dean, J, "Distilling the knowledge in a neural network," 2015.
doi: https://doi.org/10.48550/arXiv.1503.02531

[13] Chong, Z., Ma, X., Zhang, H., Yue, Y., Li, H., Wang, Z., Ouyang, W, "Monodistill: Learning spatial features for monocular 3d object detection," Proc. ICLR, 2022.
doi: https://doi.org/10.48550/arXiv.2201.10830

[14] Zhou, S., Liu, W., Hu, C., Zhou, S., Ma, C, "UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 5116-5125, 2023.
doi: https://doi.org/10.48550/arXiv.2303.15083

[15] Zhou, S., Liu, Z., Hu, C., Shi, S., Wei, Y., Zhang, J., Li, H, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 10012-10022. 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00986

[16] Loshchilov, I., Hutter, F, "Decoupled weight decay regularization," Proc. ICLR, New Orleans, Louisiana, 2019.
doi: https://doi.org/10.48550/arXiv.1711.05101

[17] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Beijbom, O, "nuScenes: A Multimodal Dataset for Autonomous Driving," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 11621-11631, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.01164

[18] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Anguelov, D, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 2446-2454, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00252

저 자 소 개

황 혜 린

- 2020년 : 충북대학교 환경생명화학과 졸업(농학사)
- 2022년 ~ 현재 : 충남대학교 컴퓨터공학과 석사과정
- ORCID : https://orcid.org/0009-0008-0848-8908
- 주관심분야 : 영상처리, 딥러닝

조 동 현

- 2019년 : KAIST 전기 및 전자공학부 졸업(공학박사)
- 2019년 ~ 현재 : 충남대학교 전자공학과 조교수
- ORCID : https://orcid.org/0000-0002-2184-921X
- 주관심분야 : 컴퓨터 비전, 딥러닝