(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0069805
(43) 공개일자 2025년05월20일

(51) 국제특허분류(Int. Cl.)
G06V 10/80 (2022.01) *G01S 17/894* (2020.01)
G06N 3/0455 (2023.01) *G06N 3/0895* (2023.01)
G06V 10/12 (2022.01) *G06V 10/82* (2022.01)
G06V 20/56 (2022.01) *G06V 20/64* (2022.01)
G06V 30/10 (2022.01)
(52) CPC특허분류
G06V 10/806 (2023.08)
G01S 17/894 (2022.01)
(21) 출원번호 10-2024-0032793
(22) 출원일자 2024년03월07일
심사청구일자 없음
(30) 우선권주장
1020230155437 2023년11월10일 대한민국(KR)

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
장수진
경기도 화성시 동탄반석로 70, 435동 1503호 (반송동, 솔빛마을신도브래뉴아파트)
김상필
서울특별시 성북구 안암로 145, 우정정보관 606호 (안암동5가)
(뒷면에 계속)
(74) 대리인
특허법인 무한

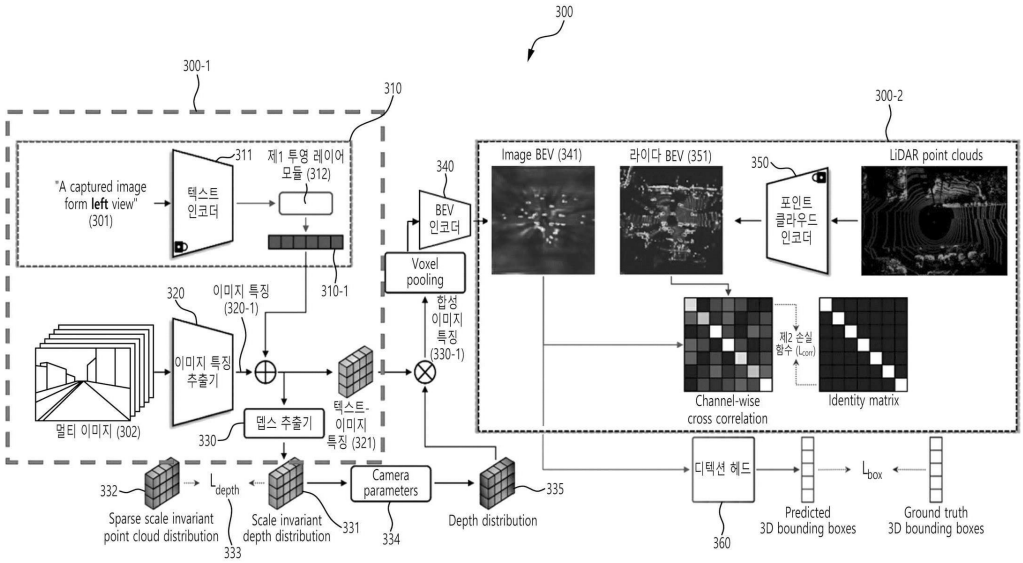
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **객체 인식 모델의 학습 방법 및 그 장치**

(57) 요 약

아래의 개시는 객체 인식 모델의 학습 방법 및 장치에 관한 것으로, 사전 학습된(pre-trained) 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하는 텍스트-가이드 학습 동작, BEV 인코더 및 포인트 클라우드 인코더를 이용하는 라이다-가이드 학습 동작 및 텍스트-가이드 학습 결과 및 라이다-가이드 학습 결과에 기초하여, 객체 인식 모델을 학습하는 동작을 포함하고, 텍스트-가이드 학습 동작은 하나 이상의 텍스트 입력 및 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하는 동작 및 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여, 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하는 동작을 포함할 수 있다.

대 표 도



(52) CPC특허분류
G06N 3/0455 (2023.01)
G06N 3/0895 (2023.01)
G06V 10/12 (2023.08)
G06V 10/82 (2022.01)
G06V 20/56 (2023.08)
G06V 20/64 (2023.08)
G06V 30/10 (2023.08)
G06T 2207/10028 (2013.01)
G06T 2210/12 (2013.01)

(72) 발명자
김진규
서울특별시 성북구 안암로 145, 고려대학교 내 정보대학 (안암동5가)
류원정
서울특별시 성북구 안암로9가길 53, 402호 (안암동5가)
이동욱
경기도 용인시 기흥구 서천동로 22, 505동 601호 (서천동, 힐스테이트서천)
장규삼
경기도 성남시 분당구 동판교로 275, 112동 703호 (삼평동, 봇들마을1단지)
지대현
경기도 화성시 동탄문화센터로 38, 414동 1001호 (반송동, 동탄솔빛마을 서해그랑블아파트)

명 세 서

청구범위

청구항 1

사전 학습된(pre-trained) 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하는 텍스트-가이드 학습 동작;

BEV 인코더 및 포인트 클라우드 인코더를 이용하는 라이다-가이드 학습 동작; 및

상기 텍스트-가이드 학습 결과 및 상기 라이다-가이드 학습 결과에 기초하여, 객체 인식 모델을 학습하는 동작

을 포함하고,

상기 텍스트-가이드 학습 동작은

하나 이상의 텍스트 입력 및 상기 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하는 동작; 및

상기 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여, 상기 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하는 동작

을 포함하는, 객체 인식 모델의 학습 방법.

청구항 2

제1항에 있어서,

상기 하나 이상의 텍스트-이미지 특징을 출력하는 동작은

상기 하나 이상의 텍스트 입력에 대해 상기 텍스트-가이드 모델에 포함된 텍스트 인코더 및 제1 투영 레이어 모듈을 통해 의미론적(semantic) 정보 인코딩을 수행하여, 하나 이상의 카메라 변형(camera-variant) 정보를 출력하는 동작; 및

상기 하나 이상의 카메라 변형 정보와 상기 이미지 특징 추출기에서 추출된 하나 이상의 이미지 특징을 더하여, 상기 텍스트-이미지 특징을 생성하는 동작

을 포함하는, 객체 인식 모델 학습 방법.

청구항 3

제1항에 있어서,

상기 라이다-가이드 학습 동작은

상기 포인트 클라우드 인코더로부터 획득한 라이다 BEV들과 상기 BEV 인코더에 기초하여 생성된 이미지 BEV들을 대조 학습하는 동작

을 포함하는, 객체 인식 모델 학습 방법.

청구항 4

제3항에 있어서,

상기 대조 학습하는 동작은

상기 라이다 BEV와 상기 이미지 BEV에 대한 교차 상관 관계를 제2 손실 함수에 기초하여 학습하는 동작

을 포함하는, 객체 인식 모델 학습 방법.

청구항 5

제3항에 있어서,

상기 객체 인식 모델을 학습하는 동작은

상기 하나 이상의 텍스트-이미지 특징 또는 상기 대조 학습 결과에 기초하여, 상기 이미지 특징 추출기, 뎁스 추출기, 상기 BEV 인코더, 디텍션 헤드 중 적어도 하나를 업데이트하는 동작

을 포함하는, 객체 인식 모델 학습 방법.

청구항 6

제5항에 있어서,

상기 뎁스 추출기는

상기 하나 이상의 텍스트-이미지 특징에 기초하여 제1 뎁스 정보를 생성하고,

상기 제1 뎁스 정보, 뎁스 추출 포인트 클라우드로부터 생성된 제2 뎁스 정보 및 뎁스 손실 함수에 기초하여 업데이트되는, 객체 인식 모델 학습 방법.

청구항 7

제6항에 있어서,

상기 BEV 인코더는

상기 뎁스 추출기를 이용하여 추출된 뎁스 특징 및 상기 텍스트-이미지 특징에 기초하여 생성된 합성 이미지 특징에 기초하여, 상기 이미지 BEV를 생성하는, 객체 인식 모델 학습 방법.

청구항 8

제1항에 있어서,

상기 사전 학습된 텍스트-가이드 모델은

텍스트 인코더, 제2 투영 레이어 모듈 및 이미지 인코더를 활용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하는 텍스트-가이드 모델 학습 방법에 의해 획득되는, 객체 인식 모델 학습 방법.

청구항 9

제8항에 있어서,

상기 텍스트-가이드 모델 학습 방법은

학습용 텍스트 입력 및 학습용 이미지 입력을 수신하고,

상기 학습용 텍스트 입력에서 상기 텍스트 인코더 및 상기 제2 투영 레이어 모듈을 이용하여 학습용 텍스트 특징-학습용 카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간에 투영하고,

상기 학습용 이미지 입력에서 상기 이미지 인코더를 이용하여 학습용 이미지 특징을 추출하여, 상기 공유 임베딩 공간에 투영하고,

미리 결정된 방법을 이용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하는 방법인, 객

체 인식 모델 학습 방법.

청구항 10

제9항에 있어서,

상기 미리 결정된 방법은

상기 학습용 텍스트 특징과 상기 학습용 이미지 특징을 상기 공유 임베딩 공간에서 대조 정렬 학습하고,

상기 대조 정렬 학습 결과 및 제1 손실 함수를 이용하여, 상기 제2 투영 레이어 모듈을 상기 제1 투영 레이어 모듈로 학습하는 방법인, 객체 인식 모델 학습 방법.

청구항 11

제10항에 있어서,

상기 제1 손실 함수는

상기 학습용 텍스트 특징에서 불분명한 기하학적 노이즈를 억제하는 카메라 분류기를 포함하는, 객체 인식 모델 학습 방법.

청구항 12

텍스트-이미지 쌍의 입력으로부터 카메라 변형 정보를 생성하는 사전 학습된 텍스트-가이드 모델;

상기 텍스트-이미지 쌍의 입력으로부터 이미지 특징을 추출하는 이미지 특징 추출기;

상기 이미지 특징에 기초하여 뎁스 정보를 추출하는 뎁스 추출기;

상기 카메라 변형 정보, 상기 이미지 특징 및 상기 뎁스 정보에 기초하여 이미지 BEV를 생성하는 이미지 BEV; 및

상기 이미지 BEV에 기초하여, 객체 인식을 수행하는 디텍션 헤드

를 포함하는, 텍스트-가이드 객체 인식 모델.

청구항 13

객체 인식 모델 학습 장치에 있어서,

인스트럭션들을 포함하는 메모리; 및

사전 학습된 텍스트-가이드 모델, 이미지 특징 추출기, BEV 인코더 및 라이다-가이드 모델 -포인트 클라우드 인코더를 포함함-, 디텍션 헤드를 구동하는 하나 이상의 프로세서

를 포함하고,

상기 인스트럭션들은 상기 프로세서에 의해 실행될 때, 상기 객체 인식 모델 학습 장치로 하여금

하나 이상의 텍스트 입력 및 상기 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하고,

상기 텍스트-가이드 모델 및 상기 이미지 특징 추출기를 이용하여, 상기 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하고,

상기 텍스트-이미지 특징 및 상기 라이다-가이드 모델의 결과에 기초하여, 객체 인식 모델을 학습하도록 하는, 객체 인식 모델 학습 장치.

청구항 14

제13항에 있어서,

상기 텍스트-가이드 모델은

상기 하나 이상의 텍스트 입력에 대해 텍스트 인코더 및 제1 투영 레이어 모듈을 통해 의미론적 정보 인코딩을 수행하여, 상기 하나 이상의 카메라 변형 정보를 생성하고,

상기 하나 이상의 텍스트-이미지 특징은

상기 하나 이상의 카메라 변형 정보와 상기 이미지 특징 추출기에서 추출된 하나 이상의 이미지 특징을 더하여 생성되는, 객체 인식 모델 학습 장치.

청구항 15

제13항에 있어서,

상기 라이다-가이드 모델은

상기 포인트 클라우드 인코더로부터 획득한 라이다 BEV들과 상기 BEV 인코더로부터 획득한 이미지 BEV들을 대조 학습하는 모델인, 객체 인식 모델 학습 장치.

청구항 16

제15항에 있어서,

상기 대조 학습은

상기 라이다 BEV와 상기 이미지 BEV에 대한 교차 상관 관계를 제2 손실 함수에 기초하여 상기 객체 인식 모델을 학습하는 방법인, 객체 인식 모델 학습 장치.

청구항 17

제15항에 있어서,

상기 프로세서는

상기 하나 이상의 텍스트-이미지 특징 또는 상기 대조 학습 결과에 기초하여, 상기 이미지 특징 추출기, 상기 뎁스 추출기, 상기 BEV 인코더, 상기 디텍션 헤드 중 적어도 하나를 업데이트하는, 객체 인식 모델 학습 장치.

청구항 18

텍스트-가이드 모델 학습 장치에 있어서,

인스트럭션들을 포함하는 메모리; 및

텍스트 인코더, 제2 투영 레이어 모듈 및 이미지 인코더를 포함하는 하나 이상의 프로세서

를 포함하고,

상기 인스트럭션들은 상기 프로세서에 의해 실행될 때, 상기 텍스트-가이드 모델 학습 장치로 하여금,

텍스트 입력 및 이미지 입력을 수신하고,

상기 텍스트 입력에서 상기 텍스트 인코더 및 상기 제2 투영 레이어 모듈을 이용하여 텍스트 특징-카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간에 투영하고,

상기 이미지 입력에서 상기 이미지 인코더를 이용하여 이미지 특징을 추출하여, 상기 공유 임베딩 공간에 투영

하고,

미리 결정된 방법을 이용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하여, 사전 학습된 텍스트-가이드 모델을 획득하도록 하는, 텍스트-가이드 모델 학습 장치.

청구항 19

제18항에 있어서,

상기 미리 결정된 방법은

상기 텍스트 특징과 상기 이미지 특징을 상기 공유 임베딩 공간에서 대조 정렬 학습하고,

상기 대조 정렬 학습 결과 및 제1 손실 함수를 이용하여, 상기 제2 투영 레이어 모듈을 상기 제1 투영 레이어 모듈로 학습하는 방법인, 객체 인식 모델 학습 방법.

청구항 20

제19항에 있어서,

상기 제1 손실 함수는

상기 학습용 텍스트 특징에서 불분명한 기하학적 노이즈를 억제하는 카메라 분류기를 포함하는, 객체 인식 모델 학습 방법.

발명의 설명

기 술 분 야

[0001] 아래의 개시는 객체 인식 모델의 학습 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 최근 다양한 센서(예: LiDAR, RADAR, Multi-View Camera)를 활용한 컴퓨터 비전 기술의 도입으로 객체 인식 및 관련 연구들이 빠르게 발전하고 있다. 3차원 객체 인식 기술은 객체의 크기, 위치 및 분류를 정확하게 인지하는데 핵심적인 역할을 할 수 있다. 특히, 고정밀 LiDAR 센서를 활용하여, 3차원에서 객체의 깊이 정보를 예측할 수 있다. 그러나, 고밀도 LiDAR 센서는 고비용을 요구하므로, LiDAR를 대체할 수 있는 기술의 개발이 필요할 수 있다. Multi-View Camera 기술은 차량의 여러 방위에 카메라들을 설치하여 다각도 이미지를 동시 다발적으로 처리하는 방식으로, 2차원 이미지를 3차원 Bird-Eye-View(BEV) 공간으로 투영하는데 효과적일 수 있다.

발명의 내용

과제의 해결 수단

[0003] 일 실시예에 따른 객체 인식 모델 학습 장치는 사전 학습된(pre-trained) 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하는 텍스트-가이드 학습 동작, BEV 인코더 및 포인트 클라우드 인코더를 이용하는 라이다-가이드 학습 동작 및 상기 텍스트-가이드 학습 결과 및 상기 라이다-가이드 학습 결과에 기초하여, 객체 인식 모델을 학습하는 동작을 포함하고, 상기 텍스트-가이드 학습 동작은 하나 이상의 텍스트 입력 및 상기 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하는 동작 및 상기 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여, 상기 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하는 동작을 포함할 수 있다.

[0004] 상기 하나 이상의 텍스트-이미지 특징을 출력하는 동작은 상기 하나 이상의 텍스트 입력에 대해 상기 텍스트-가이드 모델에 포함된 텍스트 인코더 및 제1 투영 레이어 모듈을 통해 의미론적(semantic) 정보 인코딩을 수행하

여, 하나 이상의 카메라 변형(camera-variant) 정보를 출력하는 동작 및 상기 하나 이상의 카메라 변형 정보와 상기 이미지 특징 추출기에서 추출된 하나 이상의 이미지 특징을 더하여, 상기 텍스트-이미지 특징을 생성하는 동작을 포함할 수 있다.

[0005] 상기 라이다-가이드 학습 동작은 상기 포인트 클라우드 인코더로부터 획득한 라이다 BEV들과 상기 BEV 인코더에 기초하여 생성된 이미지 BEV들을 대조 학습하는 동작을 포함할 수 있다.

[0006] 상기 대조 학습하는 동작은 상기 라이다 BEV와 상기 이미지 BEV에 대한 교차 상관 관계를 제2 손실 함수에 기초하여 학습하는 동작을 포함할 수 있다.

[0007] 상기 객체 인식 모델을 학습하는 동작은 상기 하나 이상의 텍스트-이미지 특징 또는 상기 대조 학습 결과에 기초하여, 상기 이미지 특징 추출기, 상기 뎁스 추출기, 상기 BEV 인코더, 디텍션 헤드 중 적어도 하나를 업데이트하는 동작을 포함할 수 있다.

[0008] 상기 뎁스 추출기는 상기 하나 이상의 텍스트-이미지 특징에 기초하여 제1 뎁스 정보를 생성하고, 상기 제1 뎁스 정보, 뎁스 추출 포인트 클라우드로부터 생성된 제2 뎁스 정보 및 뎁스 손실 함수에 기초하여 업데이트될 수 있다.

[0009] 상기 BEV 인코더는 상기 뎁스 추출기를 이용하여 추출된 뎁스 특징 및 상기 텍스트-이미지 특징에 기초하여 생성된 합성 이미지 특징에 기초하여, 상기 이미지 BEV를 생성할 수 있다.

[0010] 상기 사전 학습된 텍스트-가이드 모델은 텍스트 인코더, 제2 투영 레이어 모듈 및 이미지 인코더를 활용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하는 텍스트-가이드 모델 학습 방법에 의해 획득할 수 있다.

[0011] 상기 텍스트-가이드 모델 학습 방법은 학습용 텍스트 입력 및 학습용 이미지 입력을 수신하고, 상기 학습용 텍스트 입력에서 상기 텍스트 인코더 및 상기 제2 투영 레이어 모듈을 이용하여 학습용 텍스트 특징-학습용 카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간에 투영하고, 상기 학습용 이미지 입력에서 상기 이미지 인코더를 이용하여 학습용 이미지 특징을 추출하여, 상기 공유 임베딩 공간에 투영하고, 미리 결정된 방법을 이용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하는 방법일 수 있다.

[0012] 상기 미리 결정된 방법은 상기 학습용 텍스트 특징과 상기 학습용 이미지 특징을 상기 공유 임베딩 공간에서 대조 정렬 학습하고, 상기 대조 정렬 학습 결과 및 제1 손실 함수를 이용하여, 상기 제2 투영 레이어 모듈을 상기 제1 투영 레이어 모듈로 학습하는 방법일 수 있다.

[0013] 상기 제1 손실 함수는 상기 학습용 텍스트 특징에서 불분명한 기하학적 노이즈를 억제하는 카메라 분류기를 포함할 수 있다.

[0014] 일 실시예에 따른 텍스트-가이드 객체 인식 모델은 텍스트-이미지 쌍의 입력으로부터 카메라 변형 정보를 생성하는 사전 학습된 텍스트-가이드 모델, 상기 텍스트-이미지 쌍의 입력으로부터 이미지 특징을 추출하는 이미지 특징 추출기, 상기 이미지 특징에 기초하여 뎁스 정보를 추출하는 뎁스 추출기, 상기 카메라 변형 정보, 상기 이미지 특징 및 상기 뎁스 정보에 기초하여 이미지 BEV를 생성하는 이미지 BEV 및 상기 이미지 BEV에 기초하여, 객체 인식을 수행하는 디텍션 헤드를 포함할 수 있다.

[0015] 일 실시예에 따른 객체 인식 모델 학습 장치는 인스트럭션들을 포함하는 메모리 및 사전 학습된 텍스트-가이드 모델, 이미지 특징 추출기, BEV 인코더 및 라이다-가이드 모델 -포인트 클라우드 인코더를 포함함- , 디텍션 헤드를 구동하는 하나 이상의 프로세서를 포함하고, 상기 인스트럭션들은 상기 프로세서에 의해 실행될 때, 상기 객체 인식 모델 학습 장치로 하여금 하나 이상의 텍스트 입력 및 상기 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하고, 상기 텍스트-가이드 모델 및 상기 이미지 특징 추출기를 이용하여, 상기 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하고, 상기 텍스트-이미지 특징 및 상기 라이다 가이드 모델의 결과에 기초하여, 객체 인식 모델을 학습하도록 할 수 있다.

[0016] 상기 텍스트-가이드 모델은 상기 하나 이상의 텍스트 입력에 대해 텍스트 인코더 및 제1 투영 레이어 모듈을 통해 의미론적 정보 인코딩을 수행하여, 상기 하나 이상의 카메라 변형 정보를 생성하고, 상기 하나 이상의 텍스트-이미지 특징은 상기 하나 이상의 카메라 변형 정보와 상기 이미지 특징 추출기에서 추출된 하나 이상의 이미지 특징을 더하여 생성될 수 있다.

[0017] 상기 라이다-가이드 모델은 상기 포인트 클라우드 인코더로부터 획득한 라이다 BEV들과 상기 BEV 인코더로부터

획득한 이미지 BEV들을 대조 학습하는 모델일 수 있다.

[0018] 상기 대조 학습은 상기 라이다 BEV와 상기 이미지 BEV에 대한 교차 상관 관계를 제2 손실 함수에 기초하여 상기 객체 인식 모델을 학습하는 방법일 수 있다.

[0019] 상기 프로세서는 상기 하나 이상의 텍스트-이미지 특징 또는 상기 대조 학습 결과에 기초하여, 상기 이미지 특징 추출기, 상기 뎁스 추출기, 상기 BEV 인코더, 상기 디텍션 헤드 중 적어도 하나를 업데이트할 수 있다.

[0020] 일 실시예에 따른 텍스트-가이드 모델 학습 장치는 인스트럭션들을 포함하는 메모리 및 텍스트 인코더, 제2 투영 레이어 모듈 및 이미지 인코더를 포함하는 하나 이상의 프로세서를 포함하고, 상기 인스트럭션들은 상기 프로세서에 의해 실행될 때, 상기 텍스트-가이드 모델 학습 장치로 하여금, 텍스트 입력 및 이미지 입력을 수신하고, 상기 텍스트 입력에서 상기 텍스트 인코더 및 상기 제2 투영 레이어 모듈을 이용하여 텍스트 특징-카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간에 투영하고, 상기 이미지 입력에서 상기 이미지 인코더를 이용하여 이미지 특징을 추출하여, 상기 공유 임베딩 공간에 투영하고, 미리 결정된 방법을 이용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트하여, 사전 학습된 텍스트-가이드 모델을 획득하도록 할 수 있다.

[0021] 상기 미리 결정된 방법은 상기 학습용 텍스트 특징과 상기 학습용 이미지 특징을 상기 공유 임베딩 공간에서 대조 정렬 학습하고, 상기 대조 정렬 학습 결과 및 제1 손실 함수를 이용하여, 상기 제2 투영 레이어 모듈을 상기 제1 투영 레이어 모듈로 학습하는 방법일 수 있다.

[0022] 상기 제1 손실 함수는 상기 학습용 텍스트 특징에서 불분명한 기하학적 노이즈를 억제하는 카메라 분류기를 포함할 수 있다.

도면의 간단한 설명

[0024] 도 1a 내지 도 1c는 일 실시예에 따른 객체 인식 모델 학습 장치의 동작을 설명하기 위한 흐름도이다.

도 2는 일 실시예에 따른 텍스트-가이드 모델 학습 장치의 동작을 설명하기 위한 흐름도이다.

도 3은 일 실시예에 따른 객체 인식 모델 학습 장치를 개략적으로 도시한 것이다.

도 4는 일 실시예에 따른 텍스트-가이드 모델 학습 장치의 사전 학습 동작을 개략적으로 도시한 것이다.

도 5는 일 실시예에 따른 객체 인식 모델의 동작을 개략적으로 도시한 것이다.

도 6은 일 실시예에 따른 전자 장치의 블록도를 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 구현될 수 있다. 따라서, 실제 구현되는 형태는 개시된 특정 실시예로만 한정되는 것이 아니며, 본 명세서의 범위는 실시예들로 설명한 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[0026] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[0028] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함으로 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0029] 본 문서에서, "A 또는 B", "A 및 B 중 적어도 하나", "A 또는 B 중 적어도 하나", "A, B 또는 C", "A, B 및 C 중 적어도 하나", 및 "A, B, 또는 C 중 적어도 하나"와 같은 문구들 각각은 그 문구들 중 해당하는 문구에 함께 나열된 항목들 중 어느 하나, 또는 그들의 모든 가능한 조합을 포함할 수 있다.

[0030] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 해당 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0031] 실시예들은 퍼스널 컴퓨터, 랩톱 컴퓨터, 태블릿 컴퓨터, 스마트 폰, 텔레비전, 스마트 가전 기기, 지능형 자동차, 키오스크, 웨어러블 장치 등 다양한 형태의 제품으로 구현될 수 있다. 이하, 실시예들을 첨부된 도면들을 참조하여 상세하게 설명한다. 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성 요소는 동일한 참조 부호를 부여하고, 이에 대한 중복되는 설명은 생략하기로 한다.

[0033] 도 1a 내지 도 1c는 일 실시예에 따른 객체 인식 모델 학습 장치의 동작을 설명하기 위한 흐름도이다.

[0034] 설명의 편의를 위해, 동작들(110 내지 130, 111, 112, 112-1, 112-2)은 도 3에 도시된 객체 인식 모델 학습 장치(300)를 사용하여 수행되는 것으로 기술된다. 그러나 이 동작들(110 내지 130, 111, 112, 112-1, 112-2)은 어떤 다른 적절한 전자 기기를 통해, 그리고 어떤 적절한 시스템 내에서도 사용될 수 있을 것이다.

[0035] 나아가, 도 1의 동작은 도시된 순서 및 방식으로 수행될 수 있지만, 도시된 실시예의 사상 및 범위를 벗어나지 않으면서 일부 동작의 순서가 변경되거나 일부 동작이 생략될 수 있다. 도 1에 도시된 다수의 동작은 병렬로 또는 동시에 수행될 수 있다.

[0036] 최근 Multi-View Camera 기반 3차원 객체 인식이 Bird-Eye-View(이하, BEV라고 한다) 기술을 통해 구현되고 있다. 특정 전자 장치(예: 자율 주행 차량 등)탑재되는 객체 인식 모델의 경우 데이터의 질적/양적 한계로 인해 특정 환경에 고착화 될 수 있다. 특히, 가시성이 떨어지는 환경(예: 악천후, 밤)이나 학습되지 않은 환경에서 객체 인식에 어려움을 겪을 수 있고, 성능이 저하될 수 있다.

[0037] BEV는 객체를 위에서 바라본 시점을 의미할 수 있다. 주로 자동차의 주변 환경을 360도에서 볼 수 있게 해주는 카메라 시스템이나, 3D 컴퓨터 비전에서 객체나 장면을 위에서 내려다보는 시각적 표현에 사용될 수 있다.

[0038] 이러한, 객체 인식 모델의 성능 저하를 방지하기 위해, Unsupervised Domain Adaptation(UDA) 기술이 활용될 수 있다. UDA는 학습 과정에서, Domain Generalization 이나 Domain Adaptation 기술을 이용하여 레이블되지 않은 데이터 분포에서 성능을 확보할 수 있다. 또한, LiDAR를 활용한 객체 인식 모델에서 데이터 증강 또는 도메인 정렬을 활용한 UDA 방법론을 활용하여, 레이블되지 않은 벤치 마크 데이터 셋에서 안정적으로 3차원 객체를 예측할 수 있으나, 카메라에도 이러한 기술이 적용될 필요가 있다. 카메라의 경우, LiDAR 센서와 달리 카메라 내/외적 영향으로 인해, 뎁스 정보 오차를 발생시킬 수 있다. 이러한, 뎁스 정보 오차를 해결하기 위해, 텍스트-가이드 및 LiDAR-가이드를 이용한 객체 인식 모델의 학습 방법이 사용될 수 있다.

[0039] 이하, 텍스트-가이드 및 LiDAR-가이드를 이용한 객체 인식 모델 학습 방법을 설명한다.

[0040] 도 1의 동작들(110 내지 140)은 도 3을 참조하여 함께 설명될 수 있다.

[0041] 도 3을 참조하면, 객체 인식 모델 학습 장치(300)는 카메라 파라미터 차이로 발생하는 뎁스 정보 오차 문제를 해결하기 위해 텍스트의 일반화된 의미론적(semantic) 정보를 객체 인식 모델에 포함된 네트워크(예: 텍스트-가이드 모델(310), 이미지 특징 추출기(320), 뎁스 추출기(330), BEV 인코더(340), 디텍션 헤드(360))에 제공하여 다양한 환경에 적응한 객체 인식 모델을 학습하는 텍스트-가이드 학습(300-1)을 수행할 수 있다. 또한, 객체 인식 모델 학습 장치(300)는 BEV 인코더 및 포인트 클라우드 인코더(350)를 이용하여 이미지 BEV(341)와 라이다 BEV(351) 간의 불분명한 뎁스를 개선하는 대조 학습을 수행하는 라이다-가이드 모델을 이용하여 라이다-가이드 학습(300-2)을 수행할 수 있다.

[0042] 동작(110)에서, 객체 인식 모델 학습 장치(300)는 사전 학습된 텍스트-가이드 모델(310) 및 이미지 특징 추출기(320)를 이용하는 텍스트-가이드 학습(300-1)을 수행할 수 있다. 이를 위해, 텍스트-가이드 모델(310)은 텍스트 인코더(311) 및 제1 투영 레이어(projection layer) 모듈(312)을 포함할 수 있다.

[0043] 투영 레이어 모듈이란, Linear Projection 모듈로, MLP(Multi-Layer Perceptional)을 활용하여 구성될 수 있다. 예를 들어, 텍스트 인코더(311)는 CLIP 텍스트 인코더일 수 있고, CLIP 텍스트 인코더는 시각 데이터 표현과 직접 비교할 수 있는 텍스트 표현을 이해하고 텍스트를 생성할 수 있다.

[0044] 텍스트 인코더는 의미론적 정보 인코딩을 수행하여, 카메라 변형 정보를 생성할 수 있다. 의미론적 정보 인코딩(Semantic Information encoding)은 데이터나 신호 내의 의미를 포착하고 표현하는 방법을 의미할 수 있다. 의미론적 정보 인코딩은 주로 자연어 처리(Natural Language Processing, NLP), 이미지 인식 등에서 사용될 수 있다. 컴퓨터 시스템은 의미론적 정보 인코딩을 통해, 단어, 문장, 이미지 등의 복잡한 데이터에서 의미를 추출하고 이해할 수 있다. 텍스트 인코더는 텍스트 데이터로부터 단어의 벡터 표현을 학습하여 의미론적 정보 인코딩을 수행하는 인공 신경망 모델을 포함할 수 있다. 텍스트 인코더는 단어들 사이의 의미적 관계를 포착하여, 각 단어를 고차원 공간에서의 벡터로 변환할 수 있다. 변환된 벡터들은 단어 간의 유사성과 관계를 수학적으로 표현할 수 있다. 이를 통해, 텍스트들 간에 대수학적 연산과 유사한 연산(예: 'King' - 'Man' + 'Woman' = 'Queen')이 가능할 수 있다.

[0045] 이미지 특징 추출기(320)는 인공 신경망 모델을 이용하여 구현될 수 있다. 이미지 특징 추출기(320)는 이미지의 공간적 계층 구조를 학습하여 특징을 추출하고, 분류, 감지, 분할 등의 다양한 작업에 사용할 수 있다.

[0046] 텍스트-가이드 모델(310)은 시각 데이터 표현과 직접 비교할 수 있는 텍스트 표현을 이해하고 텍스트 특징(310-1)을 생성할 수 있다. 텍스트-가이드 모델은 보다 확실한 카메라 변형 정보를 임베딩하기 위해 대조학습을 활용하여 사전 학습(pre-trained)될 수 있다(예: 후술하는 수학식 1 및 수학식 2의 $Lcont$). 추가적으로, 명시적인 카메라 인지를 위하여 카메라 분류기(예: 후술하는 수학식 3의 Lce)를 활용할 수 있다.

[0047] 동작(111)에서, 객체 인식 모델 학습 장치(300)는 하나 이상의 텍스트(301) 입력 및 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력(302)을 수신할 수 있다. 객체 인식 모델 학습 장치(300)는 텍스트-이미지 쌍의 입력(예: 하나 이상의 텍스트(301) 및 멀티 이미지(302))을 수신할 수 있다.

[0048] 동작(112)에서, 객체 인식 모델 학습 장치(300)는 텍스트-가이드 모델(310) 및 이미지 특징 추출기(320)를 이용하여, 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징(321)을 출력할 수 있다.

[0049] 텍스트-이미지 쌍의 입력이란, 하나 이상의 카메라로부터 수신한 이미지들과 텍스트를 매칭한 입력을 의미할 수 있다. 객체 인식 모델 학습 장치(300)는 전방위 LiDAR 시스템을 모사한 하나 이상의 카메라로부터 이미지를 수신하게되면, 각각의 이미지는 카메라 정보를 포함하는 텍스트 프롬프트와 쌍을 이루게 될 수 있다(예: A car image captured from camera front left view).

[0050] 이미지의 경우 이미지 특징 추출기(320)를 통과하여 이미지 특징(320-1)으로 추출되며, 텍스트의 경우 텍스트-가이드 모델을 활용하여 카메라 변형 정보를 담고 있는 의미론적 정보를 포함하는 텍스트 특징(310-1)(또는, 카메라 변형 정보)으로 추출될 수 있다.

[0051] 동작(112-1)에서, 객체 인식 모델 학습 장치(300)는 하나 이상의 텍스트(301) 입력에 대해 텍스트-가이드 모델(310)에 포함된 텍스트 인코더(311) 및 제1 투영 레이어 모듈(312)을 통해 의미론적 정보 인코딩을 수행하여, 하나 이상의 카메라 변형 정보(또는, 텍스트 특징(310-1))을 출력할 수 있다.

[0052] 동작(112-2)에서, 하나 이상의 카메라 변형 정보와 이미지 특징 추출기(320)에서 추출된 하나 이상의 이미지 특징(320-1)을 더하여, 텍스트-이미지 특징(321)들을 생성할 수 있다.

[0053] 동작(120)에서, 객체 인식 모델 학습 장치(300)은 BEC 인코더(340) 및 포인트 클라우드 인코더(350)를 이용하는 라이다(LiDAR)-가이드 학습 동작(300-2)을 할 수 있다.

[0054] 라이다-가이드 학습 동작(300-2)는 라이다-가이드 모델을 이용하여 수행될 수 있다. 일 실시예에 따른 라이다-가이드 모델은 포인트 클라우드 인코더(350)로부터 획득한 라이다 BEV(351)들과 BEV 인코더(340)에 기초하여 생성된 이미지 BEV(341)들을 대조 학습할 수 있다. 대조 학습에서, 라이다-가이드 모델은 라이다 BEV(351)와 이미지 BEV(341)에 대한 교차 상관 관계를 제2 손실 함수에 기초하여 학습할 수 있다.

[0055] 라이다 포인트 클라우드에 포함된 정밀한 기하학적 정보는 풍부한 시맨틱 RGB 이미지를 보완할 수 있다. 따라서, 다음 수학식 1과 같이 포인트 클라우드 인코더(350)를 통해 생성된 라이다 BEV(351)와 BEV 인코더(340)를 통해 생성된 이미지 BEV(341) 간의 관계를 표현할 수 있다.

[0056] 이미지 BEV(341)는 BEV 인코더(340)가 합성 이미지 특징(330-1)으로부터 추출할 수 있다. 이 때, 합성 이미지 특징(330-1)은 텍스트-이미지 특징(321)에 뎁스 추출기(330)에서 추출된 뎁스 특징(또는, Depth distribution)(335)를 합성하여 생성될 수 있다. 또한, 이미지 BEV(341) 생성 전에 복셀 풀링(voxel pooling)을 수행할 수 있다. 복셀 풀링은 3D 딥러닝, 특히 포인트 클라우드와 같은 3D 데이터 처리에 사용될 수 있다. 복셀 풀링은 고해상도 3D 공간의 특징을 저해상도로 집계하여 중요한 공간 계층을 유지하면서 계산 복잡

성을 줄일 수 있다. 복셀 풀링은 입력 볼륨을 다운샘플링하고 정보의 양을 단순화하며 변환 불변성(translation invariance)을 어느 정도 달성하는 데 도움이 될 수 있다.

[0057] 일 실시예에 따른 뎁스 추출기(330)는 하나 이상의 텍스트-이미지 특징(321)에 기초하여 제1 뎁스 정보(331)(또는, Scale invariant depth distribution)를 생성하고, 제1 뎁스 정보(331) 및 뎁스 추출 포인트 클라우드로부터 생성된 제2 뎁스 정보(332) 및 뎁스 손실 함수(*Ldepth)*(333)에 기초하여 업데이트 될 수 있다. 이후, 뎁스 추출기(330)는 카메라 파라미터들(334)에 기초하여 뎁스 특징(335)를 추출할 수 있다.

[0058] 즉, BEV 인코더(340)는 뎁스 추출기(330)를 이용하여 추출된 뎁스 특징(335)에 기초하여 생성된 합성 이미지 특징(330-1)에 기초하여, 이미지 BEV(341)를 생성할 수 있다.

수학식 1

$$\mathcal{L}_{feat} = \frac{1}{N}\sum^{N}(BEV_{img} - BEV_{LiDAR})^2$$

[0059]

[0060] 수학식 1에서, *BEV*img와 *BEV*는 BEV 인코더(340)이 합성 이미지 특징(330-1)을 입력받아 생성한 이미지 BEV(341)과 포인트 클라우드 인코더(350)으로부터 생성된 라이다 BEV(351)일 수 있다. *L*은 RGB 기반 BEV 특징을 그리드 방식으로 표현할 수 있으나, L1/L2 distance 기반 특징 추출 기법은 불확실한 모달 정보로 부터 발생하는 모달 특화 오류(Modal-specific errors)를 전달하여 학습에 부정적인 영향을 미칠 수 있다.

[0061] 따라서, 라이다-가이드 모델은 다음 수학식 3 및 수학식 4와 같은 제2 손실 함수(*L*corr)를 사용하여, 교차 모달 중복성 정규화(Cross-modal Redundancy Regularization)를 수행할 수 있다.

수학식 2

$$\mathcal{L}_{corr} \triangleq \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}{}^2$$

[0062]

수학식 3

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2}\sqrt{\sum_b (z^B_{b,j})^2}},$$

[0063]

[0064] 수학식 2 및 수학식 3에서, *Cii*를 포함하는 항은 대각 요소를 완전한 상관 관계(즉, 1)로 만드는 불변성 항(invariance term)이고, *Cij*를 포함하는 항은 대각선 이외의 요소를 불완전한 상관 관계(즉, 0)으로 만드는 중복성 감소 항(redundancy reduction term)일 수 있다(도 3의 Channel-wise cross correlation 및 Identity matrix 참조).

[0065] 여기서, C는 -1(즉, 완벽한 반상관관계)과 1(즉, 완벽한 상관관계) 사이의 정사각형 행렬을 나타낼 수 있다. 특히, 불변성 항은 이미지 BEV(341)에서 전경과 배경 사이의 모호한 경계에 대한 견고성(robustness)을 향상시킬 수 있다. 중복 감소 항은 부적절한 교차 모달 정보를 억제할 수 있다. 결과적으로 Lcorr는 공유 임베딩 공간에서 멀티 모달리티 간 최적화된 전이 학습을 통해 다양한 입력 이미지로부터 안정적인 BEV 이미지 생성을 유도할 수 있다.

[0066] 객체 인식 모델 학습 장치(300)는 전술한 교차 모달 중복성 정규화(Cross-modal Redundancy Regularization)을 수행하는 라이다-가이드 모델을 통해 라이다-가이드 학습 동작을 수행하여, 이미지 특징 추출기(320), 뎁스 추출기(330), BEV 인코더(340) 또는 디텍션 헤드를 학습시키고, 합성 이미지 특징(330-1)으로부터 생성된 이미지 BEV(341)를 더욱 명확하게 생성할 수 있다.

[0067] 일 실시예에 따른 디텍션 헤드(360)는 이미지 BEV(341)를 입력으로 하여 예상 3D 바운딩 박스들(predicted 3D bounding boxes)(또는, 예상 객체 인식 결과)를 추출하고, 추출된 예상 3D 바운딩 박스들과 그라운드 트루스 3D 바운딩 박스들(Ground truth 3D bounding boxes)(예: 레이블링된 실제 객체 인식 결과)및 디텍션 손실 함수(*Lbox*)에 기초하여, 디텍션 헤드를 업데이트할 수 있다.

[0068] 동작(130)에서, 객체 인식 모델 학습 장치(300)는 텍스트-가이드 학습(300-1) 및 라이다-가이드 학습(300-2)에 기초하여, 객체 인식 모델을 학습할 수 있다. 결론적으로, 전술한 바와 같이 객체 인식 모델 학습 장치(300)는 하나 이상의 텍스트-이미지 특징(321) 또는 대조 학습 결과(또는, 라이다-가이드 학습 결과)에 기초하여, 이미지 특징 추출기(320), 뎁스 추출기(330), BEV 인코더(340) 및 디텍션 헤드(360) 중 적어도 하나를 업데이트할 수 있다.

[0070] 도 2는 일 실시예에 따른 텍스트-가이드 모델 학습 장치의 동작을 설명하기 위한 흐름도이다.

[0071] 도 1a 내지 도 1c를 참조한 설명은, 도 2에도 동일하게 적용될 수 있고, 중복되는 내용은 생략될 수 있다. 또한, 설명의 편의를 위하여, 도 4의 텍스트-가이드 모델 학습 장치(400)를 함께 참조하여 설명한다.

[0072] 일 실시예에 따른 객체 인식 모델 학습 장치(300)는 텍스트-가이드 모델 학습 장치(400)를 내장하여, 객체 인식 모델 학습을 시작하기 전에 텍스트-가이드 모델(310)을 획득하거나, 별개의 텍스트-가이드 모델 학습 장치(400)로부터 사전 학습된 텍스트-가이드 모델(310)을 전달 받을 수도 있다.

[0073] 텍스트-가이드 모델 학습 장치(400)에서 사전 학습된 텍스트-가이드 모델(310)은 텍스트 입력을 텍스트의 의미론적(semantic) 의미를 캡처하는 임베딩으로 변환하여 모델이 작업별 학습 없이도 광범위한 시각 및 언어 작업을 수행할 수 있도록 할 수 있다.

[0074] 2D 시각적 임베딩과 조밀하게 정렬된 CLIP(Contrastive Language-Image Pre-training) 텍스트 인코더(예: 텍스트 인코더(311 및 411))는 의미론적 텍스트 프롬프트에서 3D 기하학적 정보를 캡처하지 못하는 경우가 있을 수 있다. 따라서, 보다 정확한 3D 기하학적 정보를 획득하기 위해, 사전 학습된 텍스트-가이드 모델(310)이 필요할 수 있다

[0075] 도 4를 함께 참조하면, 텍스트-가이드 모델 학습 장치(400)는 사전 학습된(pre-trained) 텍스트-가이드 모델(310)을 획득하기 위해, 텍스트 인코더(411)(예: 도 3의 텍스트 인코더(311)) 및 이미지 인코더(420)를 활용하여, 제2 투영 레이어 모듈(412)을 사전 학습(또는, Fine-tunning)할 수 있다.

[0076] 보다 구체적으로, 사전 학습된 텍스트-가이드 모델(310)은 텍스트 인코더(411), 제2 투영 레이어 모듈(412) 및 이미지 인코더(420)를 활용하여, 제2 투영 레이어 모듈(412)을 제1 투영 레이어 모듈(312)로 업데이트하는 텍스트-가이드 모델 학습 방법에 의해 획득될 수 있다.

[0077] 동작(210)에서, 텍스트-가이드 모델 학습 장치(400)는 학습용 텍스트 입력(401) 및 학습용 이미지 입력(402)를 수신할 수 있다. 이 때, 학습용 텍스트 입력(401) 및 학습용 이미지 입력(402)은 학습용 텍스트-이미지 쌍 입력일 수 있다.

[0078] 동작(220)에서, 텍스트-가이드 모델 학습 장치(400)는 학습용 텍스트 입력(401)에서 텍스트 인코더(411) 및 제2 투영 레이어 모듈(412)를 이용하여 학습용 텍스트 특징(410-1)(학습용 카메라 변형 정보를 포함함)을 추출하여, 공유 임베딩 공간(430)에 투영할 수 있다. 제2 투영 레이어 모듈(412)은 제1 투영 레이어 모듈로 업데이트(또는, 튜닝)되기 이전의 투영 레이어 모듈일 수 있다.

[0079] 텍스트-가이드 모델 학습 장치(400)는 각각의 텍스트-이미지 쌍(예: 학습용 텍스트 입력(401) 및 학습용 이미지 입력(402))으로부터 학습용 텍스트 특징(410-1)(또는, 학습용 카메라 변형 정보) 및 학습용 이미지 특징(420-1)를 인코딩할 수 있다. 여기서, 텍스트-이미지 쌍의 입력은 I(이미지) = {i1,i2,…,in}, T(텍스트) = {t1,t2,…,tn}로 표현될 수 있다.

[0080] 동작(230)에서, 텍스트-가이드 모델 학습 장치(400)는 학습용 이미지 입력(402)에서, 이미지 인코더를 이용하여

학습용 이미지 특징(420-1)을 추출하여, 공유 임베딩 공간(430)으로 투영할 수 있다.

[0081] 텍스트-가이드 모델 학습 장치(400)는 입력 I 및 T로부터, V(예: 이미지 인코더(420)), 텍스트 인코더(411) 및 제2 투영 레이어 모듈(412)을 활용하여, 공유 임베딩 공간(430)으로 각각의 특징들을 투영할 수 있다. 이 후, 대조 정렬 학습과 카메라 분류기(예: 후술하는 수학식 4 내지 수학식 6)를 통하여 제2 투영 레이어 모듈(412)만 gradient를 업데이트 할 수 있다.

[0082] 동작(240)에서, 텍스트-가이드 모델 학습 장치(400)는 미리 결정된 방법을 이용하여, 제2 투영 레이어 모듈(412)을 제1 투영 레이어 모듈(312)로 업데이트할 수 있다. 미리 결정된 방법은 학습용 텍스트 특징(401)과 학습용 이미지 특징(402)을 공유 임베딩 공간(430)에서 대조 정렬 학습하고, 대조 정렬 학습 결과 및 제1 손실 함수를 이용하여, 제2 투영 레이어 모듈(412)을 제1 투영 레이어 모듈(312)로 학습하는 방법일 수 있다. 제1 손실 함수는 학습용 텍스트 특징(410-1)에서 불분명한 기하학적 노이즈를 억제하는 카메라 분류기를 포함할 수 있다.

[0083] 이하, 텍스트-가이드 모델 학습 장치의 동작을 수식들과 함께 상세히 설명한다.

[0084] 텍스트-가이드 모델 학습 장치(400)는 텍스트-가이드 모델(310)의 획득을 위해, 텍스트 인코더(411) 및 제2 투영 레이어 모듈(412)을 거쳐 텍스트 특징(*zt*)(410-1)을 추출할 수 있다. 텍스트-가이드 모델 학습 장치(400)는 이미지 인코더(420)를 통해 이미지 특징(*zi*)(420-1)을 추출할 수 있다. 이 후, 텍스트-가이드 모델 학습 장치(400)는 다음 수학식 4 및 수학식 5의 *Lcont*에 따라 추출된 텍스트 특징(*zt*)(410-1)과 추출된 이미지 특징(*zi*)(420-1)을 , 공유 임베딩 공간(430)에 투영하고, 추출된 텍스트 특징(*zt*)(410-1)과 추출된 이미지 특징(*zi*)(420-1) 사이의 대조 정렬(contrastively align)을 수행할 수 있다.

수학식 4

$$\mathcal{L}_{\text{cont}}^{z_t \leftrightarrow z_i} = \frac{1}{N}\sum_{i=1}^{N}(l_{sim}^{z_t \to z_i} + l_{sim}^{z_i \to z_t})$$

[0085]

수학식 5

$$l_{sim}^{a \to b} = -\log \frac{\exp(\langle \mathbf{a}_i, \mathbf{b}_i \rangle)/\tau}{\sum_j \exp(\langle \mathbf{a}_i, \mathbf{b}_j \rangle)/\tau}$$

[0086]

[0087] 수학식 4에서, $\langle\rangle$는 cosine similarity를 의미하고, τ는 temperature 파라미터를 의미한다. 더해서, 텍스트-가이드 모델 학습 장치(400)는 카메라 변형 정보를 보다 명시적으로 학습하기 위해, 다음 수학식 6과 같은 카메라 분류기를 활용할 수 있다.

수학식 6

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathcal{T}_b(z_t)log(\mathcal{T}_b(z_t))$$

[0088]

[0089] 수학식 6에서, $C \in \mathbb{R}^{N \times N_I}$ 는 Multi-View deployment 레이블을 의미한다. 카메라 분류기는 불분명한 기하학적 노이즈를 억제하고 카메라 변형 정보에 대한 학습에 도움을 줄 수 있다. 텍스트-가이드 모델 학습 장치(400)는 다음 수학식 7와 같은 제1 손실 함수를 통해 제2 투영 레이어 모듈(412)를 최적화 할 수 있다.

수학식 7

$$\mathcal{L}_{sem} = \lambda_{cont}\mathcal{L}_{\text{cont}} + \lambda_{CE}\mathcal{L}_{CE}$$

[0090]

[0091] λ_{cont} 및 λ_{CE} 는 텍스트 가이드 정보를 전환하기 위한 하이퍼 파라미터들이다. 텍스트 인코더(411) 및 이미지 인코더(420)는 학습되지 않고, 제2 투영 레이어 모듈(412)만 학습될 수 있다. 학습이 완료된 제2 투영 레이어 모듈(412)은 사전 학습(또는, 튜닝)된 제1 투영 레이어 모듈(312)일 수 있다. 따라서, 텍스트-가이드 모델 학습 장치(400)는 텍스트 인코더(311 또는 411) 및 제1 투영 레이어 모듈(312)을 포함하는 텍스트-가이드 모델(310)을 획득할 수 있다.

[0092] 사전 학습된 텍스트-가이드 모델(310)은 카메라 변형(camera-variant) 정보를 출력할 수 있다. 카메라 변형 정보는 카메라 정보를 포함하고 있는 텍스트로부터 추출될 수 있다. 특히, 텍스트-가이드 모델(310)에서 출력된 카메라 변형 정보는 카메라 별 특정 scene의 정보를 이미지 특징에 임베딩될 수 있다.

[0093] 예를 들어, 카메라 변형 정보를 포함하고 있는 텍스트로 “t1 = A car image captured from camera front view”를 가정할 수 있다. 여기서, “camera front view”가 카메라 변형 정보라고 할 수 있다. 텍스트-가이드 모델(310)은 pre-train을 통해 이러한 카메라 변형 정보를 출력할 수 있도록 텍스트-가이드 모델 학습 장치(400)를 통해 학습될 수 있다.

[0094] 이하에서, 전술한 바와 같이 사전 학습된 텍스트-가이드 모델(310)을 활용한 객체 인식 모델 학습 장치의 동작을 설명한다.

[0095] 도 3을 함께 참조하면, 일 실시예에 따른 객체 인식 모델 학습 장치(300)는 텍스트-가이드 모델(310)을 활용하여 카메라 변형 정보(또는, 텍스트 특징(310-1))를 출력하고, 카메라 변형 정보와 이미지 특징 추출기(320)로부터 출력된 이미지 특징(320-1)을 더하여 통해 객체 인식 모델을 학습에 이용하는 텍스트-가이드 학습 동작(300-1)을 수행할 수 있다.

[0096] 텍스트-가이드 모델(310)를 활용하여 텍스트-이미지(T-I) 쌍 입력으로 부터 공유 임베딩 공간(여기서, 공유 임베딩 공간은 도 4의 공유 임베딩 공간과 동일한 개념으로, 카메라 변형 정보와 이미지 특징을 더하기 위한 공간임)으로 특징들을 투영한다. 이 후, 텍스트 및 이미지 특징들은 Word2Vec에 따라 high-dimension distributional vector 공간에서 각각의 특징들이 의미론적 정보(또는, 카메라 변형 정보)를 임베딩하고 있다고 가정하며 후술하는 수학식 8과 같은 대수학식을 통해 반영될 수 있다.

[0097] 즉, 텍스트-가이드 모델(310)이 추출하고자 목표하는 것은 카메라 변형 정보를 담고 있는 카메라 별 의미론적 정보일 수 있다. 따라서, 텍스트-가이드 모델(310)이 카메라 변형 정보를 추출하기 위해서 학습용 텍스트-이미지 쌍을 활용하여 사전학습이 수행된 것이다. 이 후, 카메라 변형 정보를 추출할 수 있는 해당 텍스트-가이드 모델(310)을 객체 인식 모델 학습 과정에 활용하며, 후술하는 수학식 8과 같이 기존 이미지 특징들에 추출된 텍스트 정보를 더하는(summation) 과정을 통해 증강할 수 있다.

[0099] 도 3은 일 실시예에 따른 객체 인식 모델 학습 장치를 개략적으로 도시한 것이다.

[0100] 도 3의 하나 이상의 블록들 및 블록들의 조합은 특정 기능을 수행하는 특수 목적 하드웨어 기반 컴퓨터, 또는 특수 목적 하드웨어 및 컴퓨터 명령들의 조합에 의해 구현될 수 있다.

[0101] 도 1a 내지 도 2을 참조한 설명은 도 4에도 동일하게 적용될 수 있고, 중복되는 내용은 생략될 수 있다.

[0102] 일 실시예에 따른 객체 인식 모델 학습 장치(300)는 텍스트-가이드 학습 동작(300-1) 및 라이다 가이드 학습 동작(300-2)으로 객체 인식 모델을 학습할 수 있다. 객체 인식 모델 학습 장치(300)은 사전 학습된 텍스트-가이드 모델(310), 이미지 특징 추출기(320), 뎁스 추출기(330), BEV 인코더(340) 및 라이다 가이드 모델(포인트 클라우드 인코더(350)을 포함함)을 포함할 수 있다.

[0103] 객체 인식 모델 학습 장치(300)는 하나 이상의 텍스트 입력 및 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하고, 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여, 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하고, 텍스트-이미지 특징 및 라이다 가이드 모델의 결과에 기초하여, 객체 인식 모델을 학습하도록 할 수 있다. 각각의 구성요소들은 인공 신경망 모델로 구현될 수 있다.

[0104] 객체 인식 모델 학습 장치(300)는 다음 수학식 8과 같은 방법으로 이미지 특징에 의미론적 증강을 수행할 수 있다. 객체 인식 모델 학습 장치(300)는 텍스트 기반 카메라 변형 임베딩 (zt)를 활용한 의미론적(semantic) 증강을 수행할 수 있다.

수학식 8

$$\mathcal{A}(z_i, z_t) = z_i + \frac{z_t^t - z_t^s}{||z_t^t - z_t^s||_2}$$

[0105]

[0106] 수학식 8에서, $||\ ||_2$ 는 L2 distance를 의미하고, z_t^t 는 타겟 임베딩을 의미하고, z_t^s 는 소스 임베딩을 의미한다. 객체 인식 모델 학습 장치(300)는 멀티 뷰 이미지들 및 텍스트 프롬프트들을 텍스트-가이드 모델(310) 및 이미지 특징 추출기(320)에 입력하여, zi 및 zt를 공유 임베딩 공간(도 4의 공유 임베딩 공간(430)과는 개념적으로 같음)에 투영할 수 있다. 이 후, zt의 랜덤 샘플링을 통해, $\hat{z} = \mathcal{A}(z_i, z_t)$ 를 생성하여, View 트랜스포머에 입력할 수 있다.

[0107] View 트랜스포머는 기존 2차원 특징들을 환경적 정보와 더불어 3차원 BEV 공간에 투영하는 기술을 의미할 수 있다. 기존 Image Backbone을 통해 추출한 2차원 특징들을 depth estimation 네트워크(예: 뎁스 추출기(330))에 입력하여 depth distribution을 추출하며 추출된 depth distribution은 기존 2차원 특징과 out product를 수행하여 3차원 depth volume을 형성할 수 있다. 이후, 카메라 matrice를 활용하여 3차원 BEV 공간에 투영한다.

[0108] 객체 인식 모델 학습 장치(300)는 전술한 바와 같이 데이터 증강을 통해 다양한 환경 및 왜곡된 환경에서의 데이터를 학습할 수 있다.

[0109] 결론적으로, 객체 인식 모델 학습 장치(300)는 텍스트-가이드 학습 동작(300-1)을 통해, 텍스트-이미지 쌍의 입력을 공유 임베딩 공간(도 4의 공유 임베딩 공간(430)과 개념적으로 같음)에서 도메인 일반화를 수행하고, 카메라 변형 정보를 획득할 수 있다. 객체 인식 모델 학습 장치(300)는 획득한 카메라 변형 정보를 시맨틱 데이터 증강을 수행하여 다양한 데이터를 확보할 수 있다. 이렇게 확보된 다양한 데이터를 다시 이미지 특징 추출기(320), BEV 인코더(340), 뎁스 추출기(330) 및 디텍션 헤드(360)에 학습시켜 다양한 환경에서 객체 인식을 수행하는 객체 인식 모델을 획득할 수 있다.

[0110] 객체 인식 학습 장치(300)는 전술한 텍스트-가이드 학습 동작(300-1) 및 라이다-가이드 학습 동작(300-2)를 병렬적으로 또는 순차적으로 반복하여, 객체 인식 모델에 포함된 인공 신경망 모델들을 업데이트 할 수 있다.

[0111] 이하에서, 객체 인식 모델 학습 장치(300)의 동작에 대한 예시적인 학습 과정을 설명한다.

[0112] 예를 들어, 이미지가 차량의 좌측에 탑재된 카메라로부터 수신된 이미지라고 가정한다. 이 때, 텍스트-이미지 쌍의 입력은 '좌측 카메라로부터 촬영된 이미지' - '좌측 이미지' 일 수 있다. 텍스트-가이드 모델(310)는 '좌측 카메라로부터 촬영된 이미지' 라는 텍스트를 의미론적 정보 인코딩을 통하여, 좌측 카메라 변형 정보를

생성할 수 있다.

[0113] 텍스트-가이드 모델 학습 동작에서, 객체 인식 모델 학습 장치(300)는 생성된 좌측 카메라 변형 정보를 이미지 특징에 의미론적 증강할 수 있다. 이미지 특징 추출기(320)는 '좌측 이미지'로부터 좌측 이미지 특징을 추출할 수 있다. 객체 인식 모델 학습 장치(300)는 좌측 카메라 변형 정보와 좌측 이미지 특징을 더하여 이미지 특징에 의미론적 증강을 수행할 수 있다.

[0114] 텍스트-이미지 특징(321)은 뎁스 추출기(330)에 입력되어 뎁스가 추출될 수 있다. 여기서, 뎁스 추출기(330)는 증강된 데이터에 기반하여 뎁스 손실 함수를 이용하여 학습될 수 있다.

[0115] 추출된 뎁스는 다시 합성 이미지 특징(321)과 연산될 수 있고, BEV 인코더(340)는 합성 이미지 특징(330-1)에 기초하여, 이미지 BEV(341)를 생성할 수 있다. 객체 인식 학습 장치(300)는 이미지 BEV(341) 생성 전에 복셀 풀링(Voxel Pooing)을 수행할 수 있다.

[0116] 디텍션 헤드(360)는 이미지 BEV(341)에 기초하여, 객체 인식을 수행할 수 있다. 여기서, 디텍션 헤드(360)는 디텍션 손실 함수에 기초하여, 학습될 수 있다.

[0117] 라이다-가이드 학습 동작(300-2)에서, 객체 인식 모델 학습 장치(300)는 라이다 BEV(351)와 이미지 BEV(341)를 교차 상관 관계 및 제2 손실 함수를 이용하여 BEV 인코더(340), 디텍션 헤드(360), 뎁스 추출기(330), 제1 이미지 특징 추출기(320)를 업데이트 할 수 있다.

[0118] 객체 인식 모델 학습 장치(300)는 텍스트-가이드 학습 동작(300-1)과 라이다-가이드 학습 동작(300-2)의 결과들을 이용하여, 백 프로파게이션을 통해 객체 인식 모델 학습 장치(300)의 인공 신경망 모델들을 학습할 수 있다.

[0120] 도 4는 일 실시예에 따른 텍스트-가이드 모델 학습 장치의 사전 학습 동작을 개략적으로 도시한 것이다.

[0121] 도 4의 하나 이상의 블록들 및 블록들의 조합은 특정 기능을 수행하는 특수 목적 하드웨어 기반 컴퓨터, 또는 특수 목적 하드웨어 및 컴퓨터 명령들의 조합에 의해 구현될 수 있다.

[0122] 도 1a 내지 도 3을 참조한 설명은 도 4에도 동일하게 적용될 수 있고, 중복되는 내용은 생략될 수 있다.

[0123] 일 실시예에 따른 텍스트-가이드 모델 학습 장치(400)는 사전 학습된(pre-trained) 텍스트-가이드 모델(310)을 획득하기 위해, 텍스트 인코더(411)(예: 도 3의 텍스트 인코더(311)) 및 이미지 인코더(420)를 활용하여, 제2 투영 레이어 모듈(412)을 사전 학습(또는, Fine-tunning)할 수 있다. 학습용 텍스트 입력(401)은 텍스트 인코더(411) 및 제2 투영 레이어 모듈(412)을 통해 공유 임베딩 공간(430)에 학습용 텍스트 특징(410-1)으로 투영될 수 있다. 학습용 이미지 임력(402)는 이미지 인코더(420)을 통해 공유 임베딩 공간(430)에 학습용 이미지 특징(420-1)으로 투영될 수 있다. 투영된 학습용 텍스트 특징(410-1) 및 학습용 이미지 특징(420-1)은 대조 정렬 학습, 카메라 분류기 및 제1 손실 함수에 기초하여, 제2 투영 레이어 모듈(412)를 제1 투영 레이어 모듈(312)로 업데이트 할 수 있다.

[0125] 도 5는 일 실시예에 따른 객체 인식 모델의 동작을 개략적으로 도시한 것이다.

[0126] 도 5의 하나 이상의 블록들 및 블록들의 조합은 특정 기능을 수행하는 특수 목적 하드웨어 기반 컴퓨터, 또는 특수 목적 하드웨어 및 컴퓨터 명령들의 조합에 의해 구현될 수 있다.

[0127] 도 1a 내지 도 4을 참조한 설명은 도 4에도 동일하게 적용될 수 있고, 중복되는 내용은 생략될 수 있다.

[0128] 일 실시예에 따른 객체 인식 모델 학습 장치(300)로부터 학습된 객체 인식 모델(500)은 학습된 사전 학습된 텍스트-가이드 모델(510)(예: 도 3의 텍스트-가이드 모델(310), 이미지 특징 추출기(520), 뎁스 추출기(530), BEV 인코더(540) 및 디텍션 헤드(560)에 기초하여, 객체 인식을 수행할 수 있다.

[0129] 더해서, 텍스트 인코더(511)(예: 도 3의 텍스트 인코더(311) 및 도 4의 텍스트 인코더(411)) 및 제1 투영 레이어(511)을 포함하는 텍스트 가이드 모델(510)을 통해, 카메라 변형 정보를 반영하여 객체 인식에 활용할 수 있다.

[0130] 타겟 멀티 이미지(502) 중 대상 이미지가 차량의 좌측에 탑재된 카메라로부터 수신된 이미지라고 가정한다. 이때, 텍스트-이미지 쌍의 입력은 '좌측 카메라로부터 촬영된 이미지' - '좌측 이미지' 일 수 있다. 텍스트-가

이드 모델(510)은 '좌측 카메라로부터 촬영된 이미지' 라는 텍스트 입력(501)를 의미론적 정보 인코딩을 통하여, 좌측 카메라 변형 정보(510-1)를 생성할 수 있다.

[0131] 객체 인식 모델(500)은 텍스트-가이드 모델(510)을 이용하여 생성된 좌측 카메라 변형 정보(510-1) 및 이미지 특징 추출기(520)로부터 추출된 좌측 이미지 특징(521)을 추출할 수 있다. 객체 인식 모델은 데이터 증강된 좌측 카메라 변형 정보와 좌측 이미지 특징(520-1)을 더하여 텍스트-이미지 특징(521)을 생성할 수 있다.

[0132] 이미지 특징(520-1)은 뎁스 추출기(530)에 입력되어 뎁스 정보가 추출될 수 있다. 추출된 뎁스 정보는 다시 텍스트-이미지 특징(521)과 합성될 수 있고, BEV 인코더(540)는 합성된 이미지 특징에 기초하여, 이미지 BEV(541)를 생성할 수 있다. 디텍션 헤드(560)는 이미지 BEV(541)에 기초하여, 객체 인식을 수행하여 예측 결과(570)을 출력할 수 있다.

[0134] 도 6은 일 실시예에 따른 전자 장치의 블록도를 도시한 도면이다.

[0135] 도 6의 하나 이상의 블록들 및 블록들의 조합은 특정 기능을 수행하는 특수 목적 하드웨어 기반 컴퓨터, 또는 특수 목적 하드웨어 및 컴퓨터 명령들의 조합에 의해 구현될 수 있다. 도 1a 내지 도 5를 참조하여 설명한 내용은 도 6에 동일하게 적용될 수 있다. 예를 들어, 일 실시예에 따른 전자 장치(600)는 객체 인식 모델 학습 장치(300) 및 텍스트-가이드 모델 학습 장치(400) 중 적어도 어느 하나를 포함할 수 있다.

[0136] 도 6과 같이, 전자 장치(600)는 메모리(610) 및 프로세서(620)를 포함할 수 있다. 전자 장치(600)는 통신 모듈을 더 포함할 수 있고, 통신 모듈은 전송부 및 수신부를 포함할 수 있다.

[0137] 일 실시예에 따른 전자 장치(600)는 메모리(610) 및 시스템 버스 또는 다른 적절한 회로를 통해 메모리(610)와 연결된 프로세서(620)를 포함할 수 있다.

[0138] 전자 장치(600)는 메모리(610)에 프로그램 코드를 저장할 수 있다. 일 실시예에 따른 메모리(610)는 로컬 메모리 또는 하나 이상의 대용량 저장 장치들(bulk storage devices)과 같은 하나 이상의 물리적 메모리 장치들을 포함할 수 있다. 이 때, 로컬 메모리는 RAM(Random Access Memory) 또는 프로그램 코드를 실제로 실행하는 동안 일반적으로 사용되는 다른 휘발성 메모리 장치를 포함할 수 있다. 대용량 저장 장치는 HDD(Hard Disk Drive), SSD(Solid State Drive) 또는 다른 비휘발성 메모리 장치로 구현될 수 있다.

[0139] 메모리(610)에 저장된 실행 가능한 프로그램 코드가 전자 장치(600)에 의해 실행됨에 따라, 프로세서(620)에 의해 본 개시에 기재된 다양한 동작들을 수행할 수 있다. 예를 들어, 메모리(610)는 프로세서(620)가 도 1a 내지 5에 기재된 하나 이상의 동작을 수행하도록 하기 위한 프로그램 코드를 저장할 수 있다.

[0140] 구현되는 장치의 특정 유형에 따라, 전자 장치(600)는 도시된 구성 요소보다 적은 구성 요소들 또는 도 6에 도시되지 않은 추가적인 구성 요소들을 포함할 수 있다. 또한, 하나 이상의 구성 요소들은 다른 구성 요소에 포함될 수 있고, 그렇지 않으면 다른 구성 요소의 일부를 형성할 수 있다.

[0141] 일 실시예에 따른 프로세서(620)는 전자 장치(600)의 동작들을 제어하기 위한 전반적인 제어 기능들을 수행하는 하드웨어 구성이다. 예를 들어, 프로세서(620)는 전자 장치(600) 내의 메모리(610)에 저장된 프로그램들을 실행함으로써, 전자 장치(600)를 전반적으로 제어할 수 있다. 프로세서(620)는 전자 장치(600) 내에 구비된 CPU(central processing unit), GPU(graphics processing unit), AP(application processor), NPU(neural processing unit) 등으로 구현될 수 있으나, 이에 제한되지 않는다.

[0142] 프로세서(620)는 사전 학습된 텍스트-가이드 모델(310), 이미지 특징 추출기(320), BEV 인코더(340) 및 라이다-가이드 모델 -포인트 클라우드 인코더(350), 포함함-, 디텍션 헤드(360)를 구동할 수 있다. 프로세서(620)는 객체 인식 모델 학습 장치(300)로 하여금, 하나 이상의 텍스트 입력 및 하나 이상의 텍스트 입력에 대응하는 하나 이상의 이미지 입력을 수신하고, 텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여, 객체 인식 모델의 학습에 사용되는 하나 이상의 텍스트-이미지 특징을 출력하고, 텍스트-이미지 특징 및 라이다 가이드 모델의 결과에 기초하여, 객체 인식 모델을 학습하도록 할 수 있다.

[0143] 프로세서(620)는 텍스트 인코더(411), 제2 투영 레이어 모듈(412) 및 이미지 인코더(420)를 포함할 수 있다. 프로세서는 텍스트-가이드 모델 학습 장치로 하여금, 텍스트 입력 및 이미지 입력을 수신하고, 텍스트 입력에서 텍스트 인코더 및 제2 투영 레이어 모듈을 이용하여 텍스트 특징-카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간(430)에 투영하고, 이미지 입력에서 이미지 인코더를 이용하여 이미지 특징을 추출하여, 상기 공유

임베딩 공간에 투영하고, 미리 결정된 방법을 이용하여, 상기 제2 투영 레이어 모듈을 제1 투영 레이어 모듈(312)로 업데이트하여, 사전 학습된 텍스트-가이드 모델(310)을 획득하도록 할 수 있다.

[0145] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0146] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0147] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

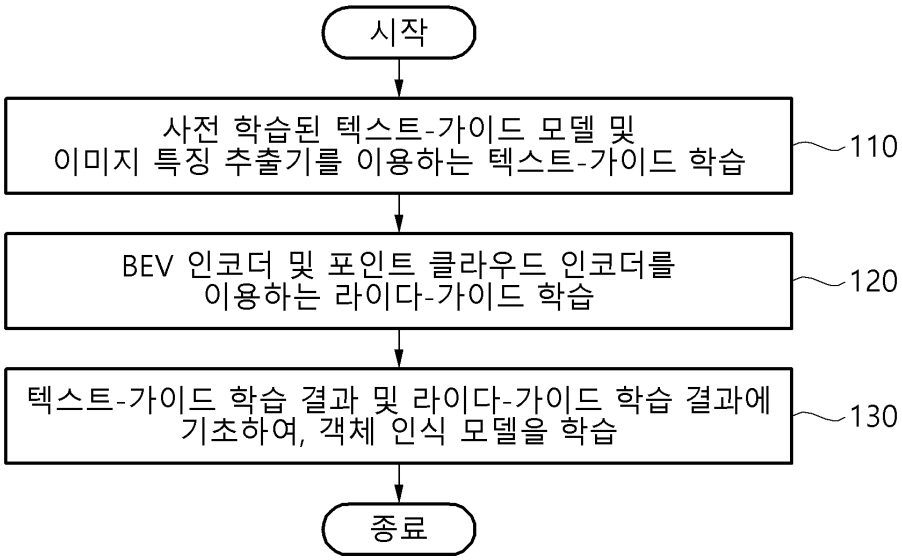
[0148] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0149] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

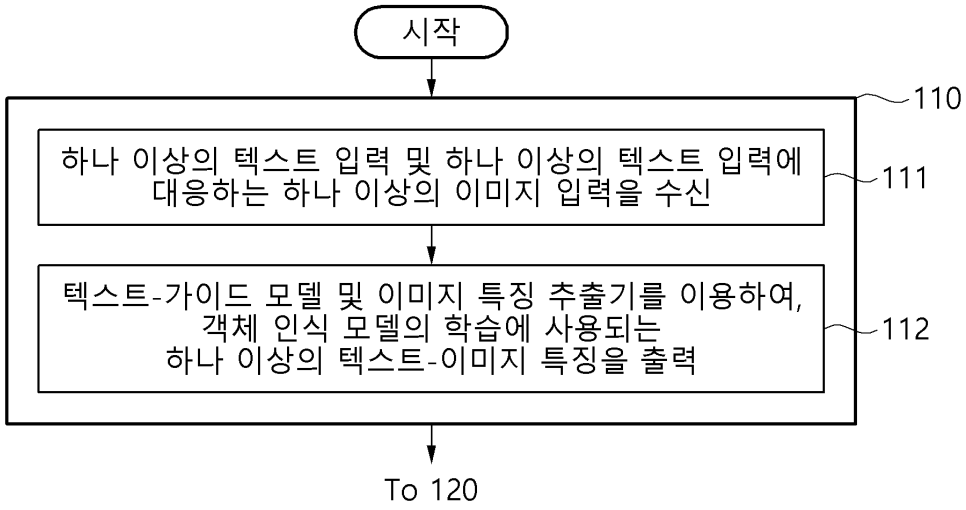
[0153]

도면

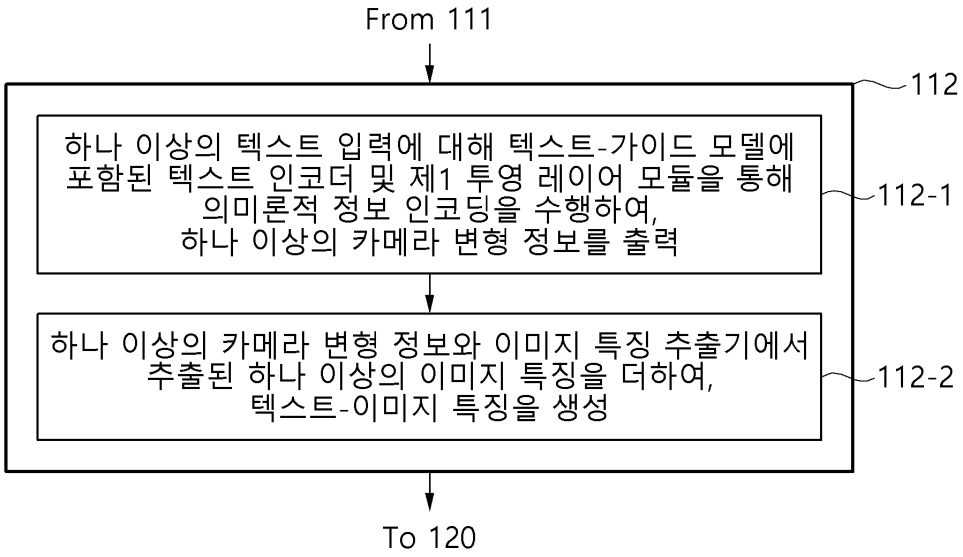
도면1a


시작
사전 학습된 텍스트-가이드 모델 및
이미지 특징 추출기를 이용하는 텍스트-가이드 학습
110
BEV 인코더 및 포인트 클라우드 인코더를
이용하는 라이다-가이드 학습
120
텍스트-가이드 학습 결과 및 라이다-가이드 학습 결과에
기초하여, 객체 인식 모델을 학습
130
종료

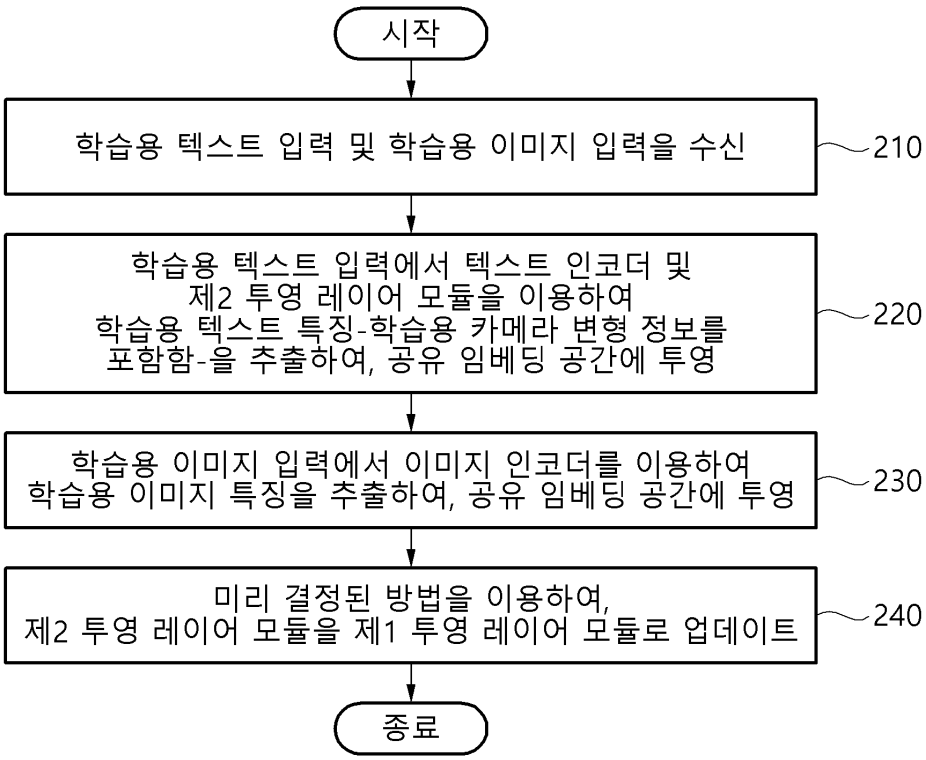
도면1b


시작
110
하나 이상의 텍스트 입력 및 하나 이상의 텍스트 입력에
대응하는 하나 이상의 이미지 입력을 수신
111
텍스트-가이드 모델 및 이미지 특징 추출기를 이용하여,
객체 인식 모델의 학습에 사용되는
하나 이상의 텍스트-이미지 특징을 출력
112
To 120

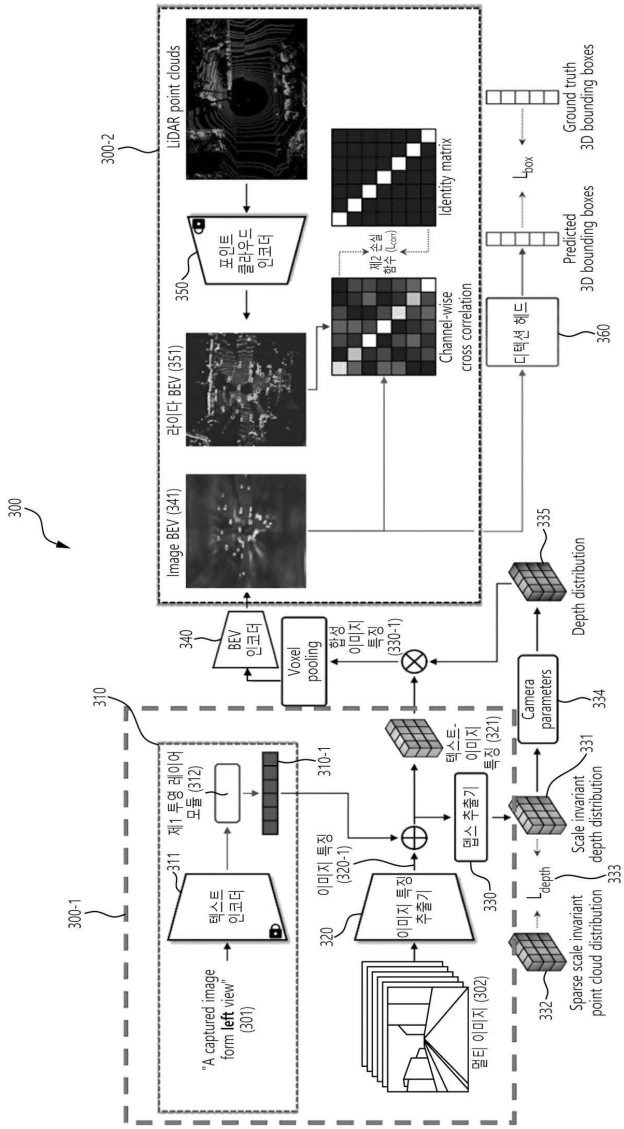
도면1c


From 111
112
하나 이상의 텍스트 입력에 대해 텍스트-가이드 모델에 포함된 텍스트 인코더 및 제1 투영 레이어 모듈을 통해 의미론적 정보 인코딩을 수행하여, 하나 이상의 카메라 변형 정보를 출력
112-1
하나 이상의 카메라 변형 정보와 이미지 특징 추출기에서 추출된 하나 이상의 이미지 특징을 더하여, 텍스트-이미지 특징을 생성
112-2
To 120

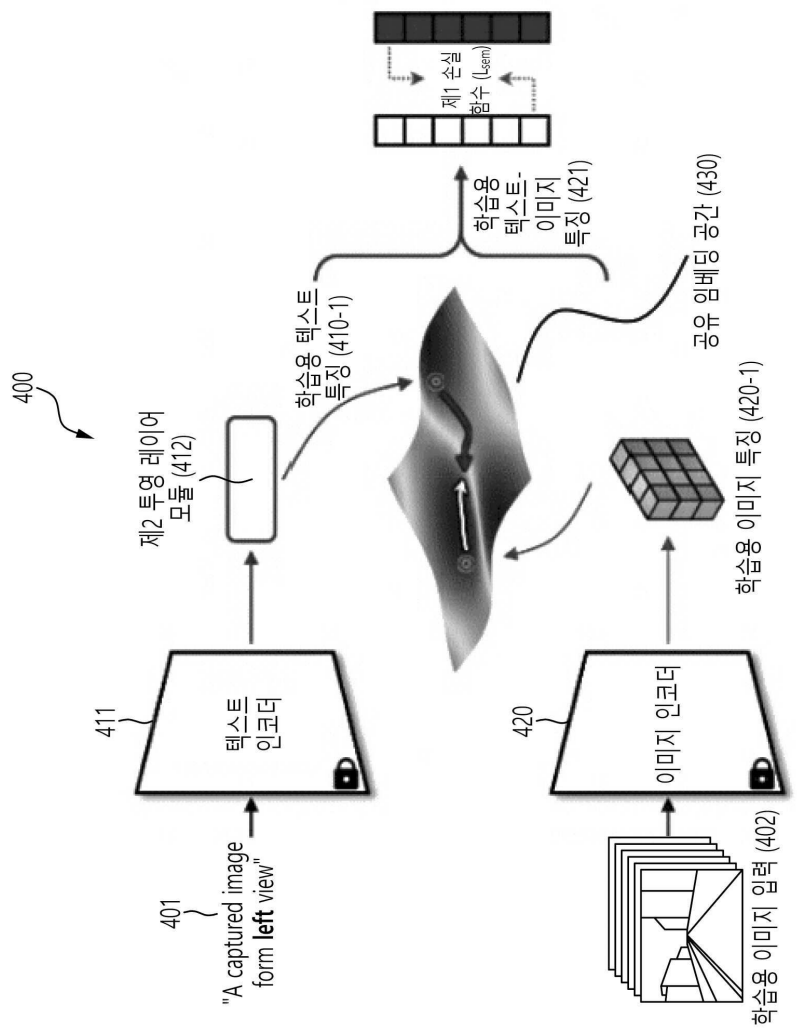
도면2


시작
학습용 텍스트 입력 및 학습용 이미지 입력을 수신
210
학습용 텍스트 입력에서 텍스트 인코더 및 제2 투영 레이어 모듈을 이용하여 학습용 텍스트 특징-학습용 카메라 변형 정보를 포함함-을 추출하여, 공유 임베딩 공간에 투영
220
학습용 이미지 입력에서 이미지 인코더를 이용하여 학습용 이미지 특징을 추출하여, 공유 임베딩 공간에 투영
230
미리 결정된 방법을 이용하여, 제2 투영 레이어 모듈을 제1 투영 레이어 모듈로 업데이트
240
종료

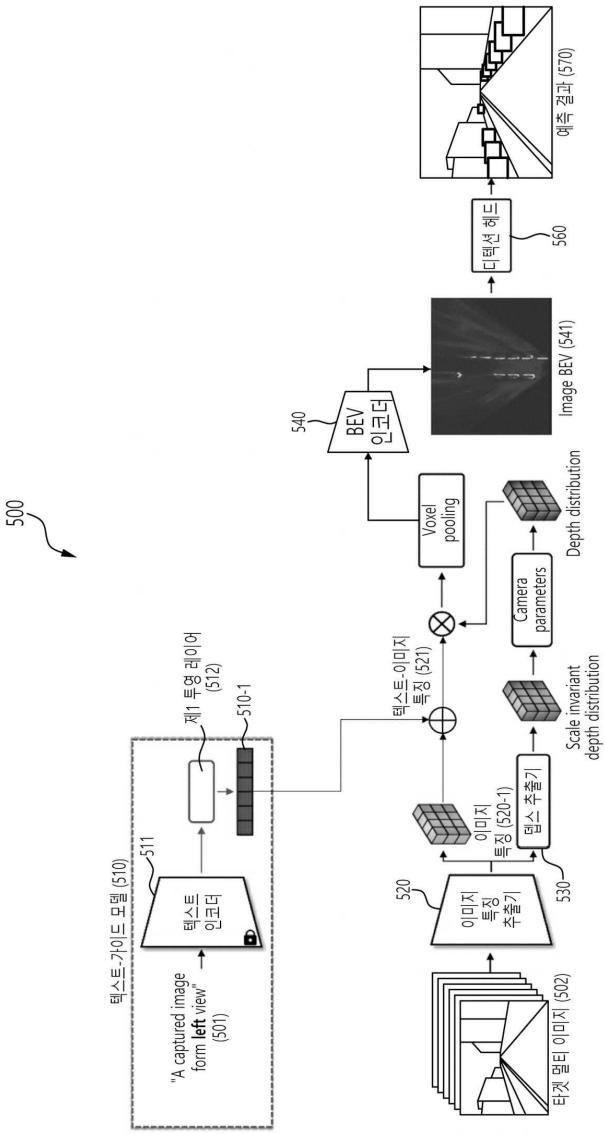
도면3


300
300-1
300-2
310
310-1
311
312
320
320-1
330
331
332
333
334
335
340
350
351
360
341
302
301
321
LiDAR point clouds
포인트 클라우드 인코더
라이다 BEV (351)
Image BEV (3341)
BEV 인코더
Voxel pooling
합성 이미지 특징 (330-1)
Identity matrix
제곱 손실 함수 (Lconf)
Channel-wise cross correlation
디텍션 헤드
Ground truth 3D bounding boxes
Predicted 3D bounding boxes
Lbox
Depth distribution
Camera parameters
텍스트 이미지 특징 (321)
뎁스 추출기
Scale invariant depth distribution
Ldepth
Sparse scale invariant point cloud distribution
이미지 특징 추출기
이미지 특징 (320-1)
멀티 이미지 (302)
제1 투영 레이어 모듈 (312)
텍스트 인코더
"A captured image form left view" (301)

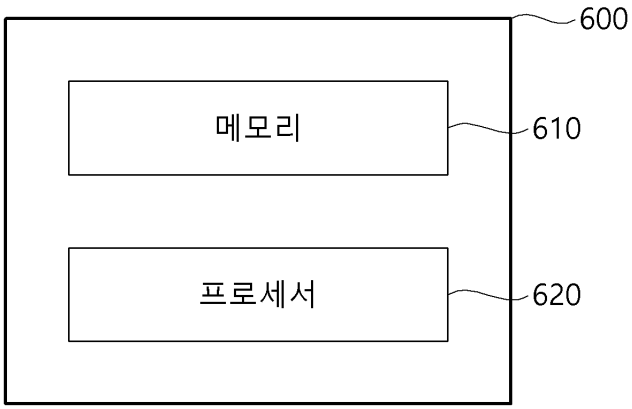
도면4


400
제2 투영 레이어 모듈 (412)
학습용 텍스트 특징 (410-1)
학습용 텍스트-이미지 특징 (421)
제1 손실 함수 (L_{sem})
공유 임베딩 공간 (430)
학습용 이미지 특징 (420-1)
411
텍스트 인코더
420
이미지 인코더
401
"A captured image form left view"
학습용 이미지 입력 (402)

도면5


500
텍스트-가이드 모델 (510)
"A captured image form left view" (501)
텍스트 인코더
511
제1 투영 레이어 (512)
510-1
텍스트-이미지 특징 (521)
타겟 멀티 이미지 (502)
이미지 특징 추출기
520
이미지 특징 (520-1)
뎁스 추출기
530
Scale invariant depth distribution
Camera parameters
Depth distribution
Voxel pooling
BEV 인코더
540
Image BEV (541)
디텍션 헤드
560
예측 결과 (570)

도면6


600
메모리
610
프로세서
620