(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0054433
(43) 공개일자 2024년04월26일

(51) 국제특허분류(Int. Cl.)
G06T 7/73 (2017.01) *B60W 40/02* (2006.01)
B60W 50/00 (2006.01) *G01S 17/89* (2020.01)
G06N 20/00 (2019.01) *G06T 5/00* (2024.01)
G06T 5/50 (2024.01) *G06T 7/11* (2017.01)
(52) CPC특허분류
G06T 7/75 (2017.01)
B60W 40/02 (2013.01)
(21) 출원번호 10-2022-0133839
(22) 출원일자 2022년10월18일
심사청구일자 없음

(71) 출원인
현대자동차주식회사
서울특별시 서초구 헌릉로 12 (양재동)
기아 주식회사
서울특별시 서초구 헌릉로 12 (양재동)
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
박형욱
서울특별시 강남구 개포로109길 9, 218동 1208호
신장호
경기도 용인시 기흥구 보정로 30, 112동 202호
(뒷면에 계속)
(74) 대리인
(유)한양특허법인

전체 청구항 수 : 총 12 항

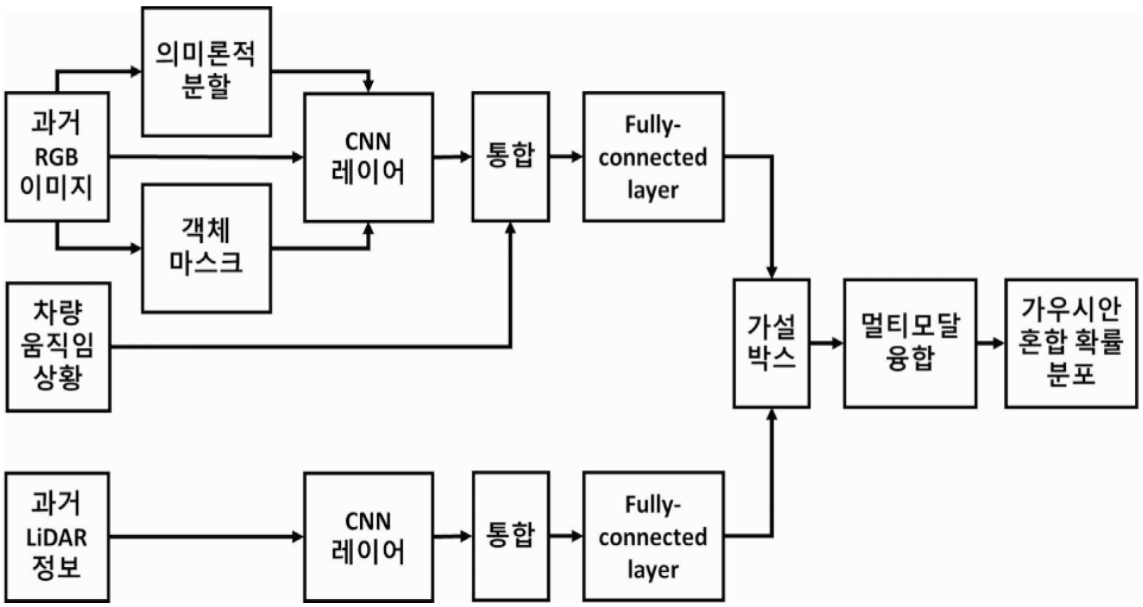
(54) 발명의 명칭 **자율주행 차량에서 객체의 미래 시점 위치 예측 방법.**

(57) 요 약

본 발명은 차량의 카메라를 통해 획득한 현재 시점 및 현재 시점 이전 복수 시점의 영상 이미지 정보를 각각 의미론적 분할(semantic segmentation) 이미지로 추출하는 단계, 상기 현재 시점 및 현재 시점 이전 복수 시점의 영상 이미지 각각에 존재하는 특정 객체(object)의 속성 및 위치 정보를 이미지화 시킨 마스크(mask) 이미지를

(뒷면에 계속)

대 표 도 - 도10



추출하는 단계, 상기 현재 시점 및 현재 시점 이전 복수 시점의 영상 이미지, 복수 개의 상기 의미론적 분할 이미지, 복수 개의 상기 마스크 이미지 및 상기 차량의 움직임 상황(ego-motion) 정보를 입력 받아, 딥 러닝(Deep Learning)을 통해 상기 객체의 미리 시점에서의 예측 위치에 대한 복수 개의 가설(hypotheses)을 도출하는 객체 위치 분포 예측 단계 및 복수 개의 상기 가설을 가우시안(Gaussian) 혼합 확률 분포로 산출하는 단계를 포함하는 자율주행 차량에서 객체의 미래 시점 위치 예측 방법.으로서, 본 발명에 의하면, 미래 예측을 위하여 차량에서 취득 가능한 다양한 정보(멀티모달)를 합성하여 미래 예측을 수행하여, 각 모달의 미래 예측에서는 현재 사건을 판단하고 판단된 현재 사건을 기준으로 미래 움직임을 예측함으로써, 최종적으로 여러 모달들의 미래 예측을 융합하여 미래 예측을 수행할 수 있다.

(52) CPC특허분류
B60W 50/0097 (2013.01)
G01S 17/89 (2022.01)
G06N 20/00 (2021.08)
G06T 5/50 (2024.01)
G06T 5/75 (2024.01)
G06T 7/11 (2017.01)
B60W 2420/403 (2013.01)
B60W 2420/408 (2024.01)
G06T 2207/20076 (2013.01)

(72) 발명자
조서영
서울특별시 동작구 상도로47길 29-1
강제원
서울특별시 마포구 서강로 53, 104동 904호
이정경
서울특별시 양천구 목동동로 350, 506동 1201호

명 세 서

청구범위

청구항 1

(a) 주행 차량의 카메라를 통해 획득한 영상 이미지를 컴퓨터가 추출하는 단계;

(b) 상기 영상 이미지를 의미론적 분할(semantic segmentation) 이미지로 추출하는 단계;

(c) 상기 영상 이미지에 존재하는 객체(object)의 속성 및 위치 정보를 이미지화 시킨 마스크(mask) 이미지를 추출하는 단계;

(d) 상기 영상 이미지, 상기 의미론적 분할 이미지 및 상기 마스크 이미지, 상기 차량의 움직임(ego-motion) 정보를 융합하는 단계;

(e) 상기 객체의 미래 시점에서의 예측 위치에 대한 복수 개의 가설(hypotheses)을 도출하는 객체 위치 분포 예측 단계;

(f) 상기 객체 위치 분포 예측 단계에 의해 도출된 복수 개의 가설에 대해 훈련 데이터를 사용해 피팅(fitting)을 수행하는 단계; 및

(g) 융합 모델(mixture model)을 생성하는 단계;를 포함하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 2

청구항 1항에 있어서,

상기 영상 이미지 정보는 상기 주행 차량의 카메라를 통해 획득한 2개 이상의 영상 이미지 정보를 추출해서 스티칭(stitching, 영상접합)한 와이드 뷰 이미지인 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 3

청구항 2항에 있어서,

상기 와이드 뷰 이미지는 RGB 2D 이미지이며,

egocentric view를 기준으로 상기 RGB 2D 이미지와 라이더 정보를 합성한 멀티 뷰를 경로예측에 사용하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 4

청구항 3항에 있어서,

상기 영상 이미지, 상기 의미론적 분할 이미지 및 상기 마스크 이미지에 의한 RGB 2D 모델과, 상기 라이더 정보에 의한 라이더 모델의 출력값을 융합하여, 상기 융합 모델(mixture model)을 생성하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 5

청구항 4항에 있어서,

상기 융합 모델(mixture model)을 이용하여 가우시안 혼합 확률 분포를 생성하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 6

청구항 5에 있어서,

상기 객체 위치 분포 예측 단계는 상기 영상 이미지에 의한 최종 벡터와 상기 라이더 정보에 의한 최종 벡터를 딥 러닝-어텐션 메커니즘(Attention Mechanism)을 이용하여 융합(합성)하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 7

청구항 3항에 있어서,

상기 복수 개의 가설과 상기 라이더 정보에 의한 라이더 모델의 출력값을 융합하여, 상기 융합 모델(mixture model)을 생성하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 8

청구항 1에 있어서,

상기 (d) 단계에서 융합되는 상기 차량의 움직임(ego-motion) 정보는 현재 시점(t) 및 미래시점(t+△t)에 대한 정보인 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 9

청구항 1항에 있어서,

상기 (d) 단계에서 융합되는 상기 영상 이미지, 상기 의미론적 분할(semantic segmentation) 이미지 및 상기 마스크(mask) 이미지는 현재 시점(t) 및 과거 복수 시점으로부터 추출되는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 10

청구항 1항에 있어서,

상기 (a) 단계이전에,

상기 객체의 위치예측 단계;를 추가로 포함하며,

상기 객체의 위치예측단계;는 상기 주행 차량의 카메라를 통해 획득한 현재 시점(t)에서 추출된 영상 이미지로부터 복수 개의 가설(hypotheses)을 도출하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 11

청구항 10항에 있어서,

상기 객체의 위치예측단계;에서 도출된 복수 개의 가설(hypotheses)은 상기 (d) 단계에서 추가로 융합되는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

청구항 12

청구항 10항에 있어서,

상기 객체의 위치예측단계;는

상기 주행 차량의 카메라를 통해 획득한 현재 시점(t)에서 추출된 영상 이미지 및 현재 시점(t)에서 추출된 상

기 의미론적 분할 이미지 및 상기 차량의 움직임(ego-motion) 정보 중 하나 이상을 포함하여, 상기 복수 개의 가설(hypotheses)을 도출하는 것을 특징으로 하는,

자율주행 차량에서 객체의 미래 시점 위치 예측 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 자율 주행 차량의 예측 운행을 제어하기 위한 객체의 미래 시점의 위치를 예측하는 방법에 관한 것이다.

배 경 기 술

[0002] 자율주행차의 자율화 범위가 점차 늘어남에 따라 돌발 상황에서 자율주행차의 안전성 요구가 증대하고 있다. 따라서, 복합적이고 다양한 돌발 상황에서 지능형 영상 신호 분석 기반 위험 방지 및 예측 기술이 필요한 상황이다. 영상 및 3차원 정보 등 다양한 센서 데이터를 이용하여 주위 환경에서의 다른 차량 및 보행자 등에 관한 선제적 예측을 수행함으로써 사고 방지를 위한 자율주행차의 효율적 대응 및 안정성 증대가 가능하므로, 자율주행차에서의 미래 객체 예측 필요성이 커지고 있다.

[0003] 종래 자율주행 차량의 미래 객체 예측을 연구한 방법들을 소개한다.

[0004] 먼저, 사전 지도(Prior map) 정보와 경로 정보를 활용한 멀티 모달은 주변 차량의 미래 경로를 예측하기 위해 사전맵(prior map)을 사용한다. 이전 연구는 주행 차선을 포함하고 교차로를 통한 유효한 경로와 같은 도로 규칙을 명시적으로 표현하는 상세한 사전 지도를 사용하여 예측을 위한 심층 신경망 모델을 제시하였다. 하지만 모든 영역에 대한 이러한 모든 도로 규칙을 명시적으로 나타내는 것은 현실적으로 불가능하기 때문에 생성자 기반의 GAN 프레임워크를 제안하였다. 판별기는 예측 궤적이 도로 규칙을 따르는지 판단하고, 생성기는 이를 따르는 궤적을 예측하게 된다.

[0005] 그리고, Egocentric view를 이용한 예측 방법이 있는데, Egocentric view란 차량의 현재 위치에서 운전자 시점에서의 영상 정보를 의미한다. 단일 시점의 RGB camera를 이용하여 Egocentric view를 취득하면 운전자 시점에서 부분적이고 협소한 시야각을 가지고 있는 문제가 있다. 또한 차량의 주행 방향을 나타내는 ego-motion으로 인한 Egocentric view에서 시간에 따른 시점 변화 문제가 있게 된다. 이러한 문제를 해결하기 위한 Egocentric view에서 다른 차량과 보행자의 미래 위치를 예측하는 방법을 설명한다. 이러한 예측 문제는 미래 시간에서 객체가 있을 수 있는 위치의 사전 확률 (prior)를 학습하여 해결한다. 지도 정보와 같은 외부의 정보를 얻는 것이 아니라 현재 이미지에서 분석한 정보를 이용하여 영상 내 탐지한 객체의 클래스가 있을 수 있는 위치의 prior를 학습한다. 또한 객체가 다음 장면에서 갑작스럽게 나타날 수 있는 위치에 대한 확률 분포를 산출하도록 학습 모델을 수립한다.(Makansi, Osama, et al. "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior." roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.)

[0006] Egocentric view에서 RGB 영상을 이용해 보행자의 다중 경로를 예측하는 방법을 개발한 연구도 있다.(Atanas Poibrenski, et al. "M2P3: Multimodal multi-pedestrian path Prediction by self-driving cars with egocentric vision." NIPS Workshop, 2020. 도 1)

[0007] 해당 연구는 RGB 영상을 비롯해 보행자의 과거 경로와 크기를 입력으로 받아 경로를 예측한다. Egocentric view에서 보행자의 과거 경로를 고려할 때, 결과로 나오는 미래 경로 분포는 다중적이다. 때문에 다중 보행자 경로 예측으로 인한 멀티 모달리티는 다대일 매핑에 있어 어려움을 겪게 된다. 이러한 문제를 CVAE (Conditional Variational Autoencoders)를 이용해 해결한다. CVAE 모델은 고차원 출력 공간의 분포를 모델링 하므로, 이를 RNN(Recurrent Neural Network) 인코더-디코더 구조와 결합하는 네트워크를 제안한다.

[0008] 또한, 경로 정보와 도로 정보를 활용한 멀티 모달 방법(Fang, Liangji, et al. "Tpnet: Trajectory proposal network for motion prediction." roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. 도 2)은 미래 경로를 결정할 때 여러 개의 경로가 존재 가능하며 자기 차량의 결정 이외에도 주위 상황 및 객체를 고려하여 경로 예측하는 방법이 제안되었다. 가능한 경로 탐색의 경우의 수를 줄이기 위해 대략적인 미래 마지막 위치를 예측하고, 이를 기반으로 가능한 미래 경로의 proposal로 hypothesis를 경로

를 생성 후 선택하게 된다.

[0009] 다음으로, 라이더(LiDAR) 정보를 활용한 것으로서, Point Pillar 구조를 통한 데이터 처리 방법이 있다.(Kawasaki, Atsushi, and Akihito Seki. "Multimodal trajectory predictions for autonomous driving without a detailed prior map." roceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021. 도 3)

[0010] Point Pillar 구조란 3차원의 포인트 클라우드 정보를 2D의 pseudo image로 변환하는 과정에서 얻은 구조를 의미한다. 기존 연구에서는 이미지의 픽셀처럼 포인트 클라우드 정보를 복셀이라는 단위로 나누어 복셀의 feature를 3D CNN의 입력으로 사용하였다. 제안 방법에서는 필러 혹은 복셀로 변환한 정보를 이용하여 연속적인 포인트 클라우드 데이터 처리하고 정보가 충분히 많은 복셀만을 계산해 각 복셀에 희소하게 분포한 포인트 클라우드 데이터의 문제를 해결한다.

[0011] 그리고, 라이더 데이터의 객체 탐색 방법(Peri, Neehar, et al. "Forecasting from LiDAR via Future Object Detection." rXiv preprint arXiv:2203.16297, 022. 도 4)은 라이더 센서를 기반으로 한 객체 감지를 수행하고 그 결과를 통해 객체의 미래 위치를 예측하기 위한 심층 신경망 구조를 개발한 연구가 있다. 현재 프레임에서의 위치를 탐지하고 미래 위치를 예측하는 방식이 아닌 미래 객체 탐지 방식으로 문제를 바라보고, 각 궤적의 시작점을 미래의 객체 위치를 예측한다. 미래와 현재 위치를 다대일 방식으로 연결하여 다중적인 미래에 대해 추론하게 된다.

[0012] 나아가, 미래 라이더 데이터 생성에 관한 논문(Deng, David, and Avideh Zakhor. "Temporal LiDAR Frame Prediction for Autonomous Driving." 020 International Conference on 3D Vision (3DV). IEEE, 2020. 도 5)에서는 이전에 주어진 미래의 LiDAR 프레임을 예측하기 위한 신경망 네트워크 모델을 제안한다. 아래 그림에서와 같이 포인트 클라우드 사이의 움직임을 예측하는 FlowNet3D 모델을 통해 라이다 데이터의 시간 사이 움직임 정보를 생성한다.

[0013] 이상의 배경기술에 기재된 사항은 발명의 배경에 대한 이해를 돕기 위한 것으로서, 이 기술이 속하는 분야에서 통상의 지식을 가진 자에게 이미 알려진 종래기술이 아닌 사항을 포함할 수 있다.

선행기술문헌

비특허문헌

[0014] (비특허문헌 0001) Makansi, Osama, et al. "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior." proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

(비특허문헌 0002) Atanas Poibrenski, et al. "M2P3: Multimodal multi-pedestrian path Prediction by self-driving cars with egocentric vision." NIPS Workshop, 2020.

(비특허문헌 0003) Fang, Liangji, et al. "Tpnet: Trajectory proposal network for motion prediction."proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

(비특허문헌 0004) Kawasaki, Atsushi, and Akihito Seki. "Multimodal trajectory predictions for autonomous driving without a detailed prior map."proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

(비특허문헌 0005) Peri, Neehar, et al. "Forecasting from LiDAR via Future Object Detection." rXiv preprint arXiv:2203.16297, 022.

(비특허문헌 0006) Deng, David, and Avideh Zakhor. "Temporal LiDAR Frame Prediction for Autonomous Driving." 020 International Conference on 3D Vision (3DV). IEEE, 2020.

발명의 내용

해결하려는 과제

[0015] 본 발명은 상술한 문제점을 해결하고자 안출된 것으로서, 본 발명은 미래 예측을 위하여 차량에서 취득 가능한 다양한 정보(멀티모달)를 합성하여 미래 예측을 수행하여, 각 모달의 미래 예측에서는 현재 사건을 판단하고 판단된 현재 사건을 기준으로 미래 움직임을 예측함으로써, 최종적으로 여러 모달들의 미래 예측을 융합하여 미래 예측을 수행하는 자율주행 차량의 멀티 모달 합성을 이용한 객체의 미래 시점 위치 예측 방법을 제공하는 데 그 목적이 있다.

과제의 해결 수단

[0016] 본 발명의 일 관점에 의한 자율주행 차량에서 객체의 미래 시점 위치 예측 방법은, (a) 주행 차량의 카메라를 통해 획득한 영상 이미지를 컴퓨터가 추출하는 단계; (b) 상기 영상 이미지를 의미론적 분할(semantic segmentation) 이미지로 추출하는 단계; (c) 상기 영상 이미지에 존재하는 객체(object)의 속성 및 위치 정보를 이미지화 시킨 마스크(mask) 이미지를 추출하는 단계; (d) 상기 영상 이미지, 상기 의미론적 분할 이미지 및 상기 마스크 이미지, 상기 차량의 움직임(ego-motion) 정보를 융합하는 단계; (e) 상기 객체의 미래 시점에서의 예측 위치에 대한 복수 개의 가설(hypotheses)을 도출하는 객체 위치 분포 예측 단계; (f) 상기 객체 위치 분포 예측 단계에 의해 도출된 복수 개의 가설에 대해 훈련 데이터를 사용해 피팅(fitting)을 수행하는 단계; 및 (g) 융합 모델(mixture model)을 생성하는 단계;를 포함한다.

[0017] 그리고, 상기 영상 이미지 정보는 상기 주행 차량의 카메라를 통해 획득한 2개 이상의 영상 이미지 정보를 추출해서 스티칭(stitching, 영상접합)한 와이드 뷰 이미지인 것을 특징으로 한다.

[0018] 또한, 상기 와이드 뷰 이미지는 RGB 2D 이미지이며, egocentric view를 기준으로 상기 RGB 2D 이미지와 라이더 정보를 합성한 멀티 뷰를 경로예측에 사용하는 것을 특징으로 한다.

[0019] 나아가, 상기 영상 이미지, 상기 의미론적 분할 이미지 및 상기 마스크 이미지에 의한 RGB 2D 모델과, 상기 라이더 정보에 의한 라이더 모델의 출력값을 융합하여, 상기 융합 모델(mixture model)을 생성하는 것을 특징으로 한다.

[0020] 더 나아가, 상기 융합 모델(mixture model)을 이용하여 가우시안 혼합 확률 분포를 생성하는 것을 특징으로 한다.

[0021] 특히, 상기 객체 위치 분포 예측 단계는 상기 영상 이미지에 의한 최종 벡터와 상기 라이더 정보에 의한 최종 벡터를 딥 러닝-어텐션 메커니즘(Attention Mechanism)을 이용하여 융합(합성)하는 것을 특징으로 한다.

[0022] 또는, 상기 복수 개의 가설과 상기 라이더 정보에 의한 라이더 모델의 출력값을 융합하여, 상기 융합 모델(mixture model)을 생성하는 것을 특징으로 한다.

[0023] 한편, 상기 (d) 단계에서 융합되는 상기 차량의 움직임(ego-motion) 정보는 현재 시점(t) 및 미래시점(t+△t)에 대한 정보인 것을 특징으로 한다.

[0024] 그리고, 상기 (d) 단계에서 융합되는 상기 영상 이미지, 상기 의미론적 분할(semantic segmentation) 이미지 및 상기 마스크(mask) 이미지는 현재 시점(t) 및 과거 복수 시점으로부터 추출되는 것을 특징으로 한다.

[0025] 또한, 상기 (a) 단계이전에, 상기 객체의 위치예측 단계;를 추가로 포함하며, 상기 객체의 위치예측단계;는 상기 주행 차량의 카메라를 통해 획득한 현재 시점(t)에서 추출된 영상 이미지로부터 복수 개의 가설(hypotheses)을 도출하는 것을 특징으로 한다.

[0026] 그리고, 상기 객체의 위치예측단계;에서 도출된 복수 개의 가설(hypotheses)은 상기 (d) 단계에서 추가로 융합되는 것을 특징으로 한다.

[0027] 여기서, 상기 객체의 위치예측단계;는 상기 주행 차량의 카메라를 통해 획득한 현재 시점(t)에서 추출된 영상 이미지 및 현재 시점(t)에서 추출된 상기 의미론적 분할 이미지 및 상기 차량의 움직임(ego-motion) 정보 중 하나 이상을 포함하여, 상기 복수 개의 가설(hypotheses)을 도출하는 것을 특징으로 한다.

발명의 효과

[0028] 본 발명에서는 1) 객체 미래 예측 네트워크를 기반으로 2) 라이더가 제공하는 멀티 뷰를 활용한 멀티 뷰 합성을 제안하고, 3) 합성된 멀티 뷰를 사용하는 객체의 미래 예측을 제안하고, 4) 영상정보와 라이더정보를 활용한 멀티모달을 설명하며 멀티모달 융합 시 단순하게 더하는 구조와 어텐션 구조의 두 가지 방법을 제안한다. 또한, 확률 분포를 이용한 최종 합성의 방법을 제안한다. 이를 통해 한정된 시야각의 영상 정보의 한계를 극복할 수

있고 제안하는 모달 간의 어텐션 구조를 통해 효과적인 모달 간 융합이 가능하게 된다.

[0029] 따라서 미래의 객체 위치를 효율적으로 예측할 수 있게 된다.

[0030] 또한, 자율주행에서의 돌발 상황에서 자율주행차의 안전성 증대 및 위험 방지 및 예측이 가능하게 된다.

[0031] 이와 같이 본 발명에 의하면 선제적 예측을 수행함으로써 자율주행차의 효율적 대응 및 안정성 증대가 가능하다.

[0032] 보다 상세하게, 시야각이 좁은 기존 1인칭 시점 카메라 영상 외에 멀티뷰 합성을 통한 넓은 시야각의 영상과 전방위 환경을 종합적으로 감지하는 센서 신호를 함께 이용하여 보다 효과적인 대응이 가능하도록 한다. 미래 자율 주행 차량에는 차량 주변의 360도를 항상 감시할 수 있는 여러 종류의 고성능 카메라 센서, 라이다, 레이더 등을 이용해 정확한 상황 감지 및 빠른 대응을 지원할 것으로, 센서 하드웨어의 발달과 전장 장치의 고도화는 지능형 소프트웨어와의 융합 및 결합으로 시너지를 일으킬 수 있을 것이다.

도면의 간단한 설명

[0033] 도 1 내지 도 5는 종래의 자율주행 차량의 미래 예측 방법에 관한 것이다.

도 6은 본 발명의 방법을 개략적으로 나타낸 것이다.

도 7은 본 발명의 객체 위치 및 위치 분포 예측 방법을 도시한 것이다.

도 8은 본 발명의 객체 위치 예측 방법을 보다 자세히 도시한 것이다.

도 9는 본 발명의 객체 위치 분포 예측 방법을 보다 자세히 도시한 것이다.

도 10은 도 7 내지 도 9를 반영한 본 발명의 방법을 개략적으로 도시한 것이다.

도 11 내지 도 13은 멀티모달 융합 과정을 나타낸 것이다.

도 14 및 도 15는 멀티 뷰 이미지를 이용한 미래 시점의 객체 위치 예측 방법을 나타낸 것이다.

도 16 및 도 17은 멀티모달 융합시 활용되는 어텐션 구조를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0034] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시 예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0035] 본 발명의 바람직한 실시 예를 설명함에 있어서, 본 발명의 요지를 불필요하게 흐릴 수 있는 공지의 기술이나 반복적인 설명은 그 설명을 줄이거나 생략하기로 한다.

[0036] 도 6은 본 발명의 방법을 개략적으로 나타낸 것이다.

[0037] 도 7은 본 발명의 객체 위치 및 위치 분포 예측 방법을 도시한 것이고, 도 8은 본 발명의 객체 위치 예측 방법을 보다 자세히 도시한 것이며, 도 9는 본 발명의 객체 위치 분포 예측 방법을 보다 자세히 도시한 것이다.

[0038] 도 10은 도 7 내지 도 9를 반영한 본 발명의 방법을 개략적으로 도시한 것이다.

[0039] 이하, 도 6 내지 도 10을 참조하여 본 발명의 일 실시예에 의한 자율주행 차량의 멀티 모달 합성을 이용한 객체의 미래 시점 위치 예측 방법을 설명하기로 한다.

[0040] 본 발명은 차량의 전방에서 취득한 2D RGB 영상 정보를 비롯해, RGB 영상에서 추출한 영상의 다양한 분석 정보, 시간에 따른 변화 정보, 라이더 정보 등을 조합하여 자동차의 1인칭 시점의 미래 시간에서 나타나는 객체의 움직임 및 등장 확률을 예측하는 방식을 보인다.

[0041] 객체의 미래 위치를 예측하는 방법으로는 전체 지도 시점 (Birds eye view) 혹은 1인칭 시점(Egocentric view)의 영상정보를 활용한 방법이 있지만, 전체지도 시점을 사용하면 처리에 필요한 데이터가 너무 방대해지는 문제가 있다. 또한 1인칭 시점의 영상정보는 한정적인 시야각을 갖고 있게 된다.

[0042] 본 발명은 이런 문제점을 극복하기 위해 RGB 2D 정보의 image stitching(영상 정합)을 하여 넓은 시야각을 활용하는 방법을 제안하고, 객체 미래 움직임 예측을 위한 넓은 시야각의 RGB 2D 정보와 라이더 정보를 융합한 멀티뷰 합성 발명을 제안한다.

[0043] 기존 방법에서는 사전 정보로 준비된 맵을 기준으로 멀티 뷰 합성 후 경로 예측을 수행하지만, 본 발명에서는 도 6에서 참조되는 바와 같이, 넓은 시야각의 1인칭 시점을 기준으로 멀티 뷰 합성 후 객체 미래 움직임 예측을 수행한다. 또한 본 발명은 라이더 정보와 RGB 2D정보를 융합할 시, 서로 다른 두 모달(modality)을 융합하는 방법을 설명한다. 기존 방법에서는 단일 모달을 이용하여 미래 위치를 예측하거나, 각각의 모달을 융합할 시 최종 예측 결과 확률을 단순하게 더하는 방법을 취하지만, 본 발명에서는 하나의 딥러닝 구조에서 어텐션 연산을 통해 라이다 및 RGB 정보를 합성해 미래 예측을 가능하게 한다.

[0044] 도 6과 같이 본 발명의 미래 예측 방법은 먼저 여러 모달 중 하나를 사용하여 미래 예측을 수행한다. 모달의 종류로 카메라를 통해 수집되는 2D RGB 영상 정보, 합성된 멀티 뷰 영상, 혹은 라이더와 레이더 정보 같은 라이더 센서 등에 의해 감지한 모든 데이터가 존재한다. 각 모달의 미래 예측에서는 현재 사건을 판단하고 판단된 현재 사건을 기준으로 미래 움직임을 예측한다. 최종적으로 여러 모달들의 미래 예측을 융합하여 미래 예측을 수행하게 된다.

[0045] 이와 같이 본 발명에서는 1) 객체 미래 예측 네트워크를 기반으로 2) 라이더가 제공하는 멀티 뷰를 활용한 멀티 뷰 합성을 제안하고, 3) 합성된 멀티 뷰를 사용하는 객체의 미래 예측을 제안하고, 4) 영상정보와 라이더정보를 활용한 멀티모달을 설명하며 멀티모달 융합 시 단순하게 더하는 구조와 어텐션 구조의 두 가지 방법을 제안한다. 또한, 확률 분포를 이용한 최종 합성의 방법을 제안한다. 이를 통해 한정된 시야각의 영상 정보의 한계를 극복할 수 있고 제안하는 모달 간의 어텐션 구조를 통해 효과적인 모달 간 융합이 가능하게 된다. 따라서 미래의 객체 위치를 효율적으로 예측할 수 있게 된다.

[0046] 도 7 내지 도 9를 통해 보다 자세히 설명하며, 도 10은 이를 모식화한 것이 다.

[0047] 본 발명에서는 주행 차량의 시점에서 취득하는 비디오(egocentric view)를 이용해 주위 차량 및 보행자의 미래 위치를 자율주행 예측 시스템 장치에 의해 예측하는 방법을 기반으로 한다.

[0048] 첫 번째로 객체 위치 예측 네트워크에서는 차량의 움직임 상황 (ego-motion)을 고려하여 미래 시점에서 배경과 객체의 속성을 고려하여 각 객체가 있을 수 있는 여럿의 후보 위치를 추론한다.

[0049] 즉, 과거 시점 영상 정보, 차량 움직임(ego-motion) 정보, 과거 시점 라이더(LiDAR) 정보를 수집하여, 이를 멀티 뷰 이미지 합성 등 일정한 처리 후 학습을 통해 멀티 뷰 내 특정 객체의 미래 시점에서의 객체 위치를 예측하는 것이다.

[0050] 그리고, 객체 위치 분포 예측 네트워크에서는 각 객체 미래 위치의 확률 분포를 예측하여 최종적으로 추정한다. 두 네트워크는 ResNet-50을 기반으로 한다. 본 발명에서는 객체 위치 분포 예측 네트워크에 멀티모달 정보를 추가적으로 활용하여 객체 미래 예측의 정확도를 향상시킨다.

[0051] 도 8의 객체 위치 예측 단계를 보다 자세히 살펴보면, 우선 주행 차량의 카메라를 통해 획득한 영상 이미지를 추출하여(S11), 추출된 RGB 영상 이미지 정보에서 의미론적 분할(semantic segmentation) 이미지를 추출한(S12) 뒤 자동차, 보행자와 같은 동적 물체를 지운 정적 의미론적 분할(static sematic segmentation) 이미지를 추정하여 추출한다(S13). 이는 배경의 속성을 파악하기 위함이며, 배경의 속성과 객체의 속성 간의 연관성을 파악해 객체(object)의 위치를 예측한다(S10). 예를 들어 자동차의 속성을 가진 객체는 도로에, 보행자의 속성을 가진 객체는 인도에 위치하게 된다. 이 과정에서 차량의 움직임 상황(ego-motion)을 자율 주행 예측 시스템 장치에 의해 수집(S14), 고려하여 객체의 미래 위치를 예측하게 되며, 일 예로 총 20개의 예측 위치에 대한 가설(hypotheses)을 도출한다(S20).

[0052] 도 9의 객체 위치 분포 예측 단계를 참조하면, 객체 위치 분포 예측 네트워크는 예를 들어 과거 0.3초 전, 0.15초 전, 그리고 현재 시점의 영상 이미지 정보 프레임을 입력으로 받아(S31) 예를 들어 미래 0.3초 뒤를 예측하게 된다. 여기서 과거 입력과 미래 출력 시간에 대한 간격은 기술 목표에 따라 조정할 수 있다. 과거 RGB 이미지 3장, 의미론적 분할(semantic segmentation) 3장(S32), 객체의 속성 정보와 객체의 위치를 이미지화 시킨 마스크(mask) 이미지 3장(S33)과 차량의 움직임 상황(ego-motion) 정보(S34), 객체 위치 예측 단계(S10)에서 출력된 20개의 가설(hypotheses)(S20)이 입력으로 들어가며, 객체 위치 분포 예측 네트워크(S30)의 결과로 20개의 신규의 가설(hypotheses)을 도출한다(S31). 이를 조정 네트워크(fitting network)에 통과시켜(S40) 최종적으로 가우시안(Gaussian) 혼합 확률 분포를 산출한다(S50).

[0053] 한편, 본 발명은 멀티 모달을 이용하여 객체 위치를 예측하는 방법을 제안하며, 도 11 내지 도 13은 멀티모달 융합 과정을 나타낸 것이며, 도 14, 도 15는 멀티 뷰 이미지를 이용한 미래 시점의 객체 위치 예측 방법을 나타

낸 것이다.

[0054] 먼저, 도 11을 참조하면, 기존의 연구에서는 전체 지도 시점을 기준으로 멀티 뷰를 합성하는 방향을 착안해왔다. 본 발명에서는 egocentric view 이미지를 활용하기 위해 2개 이상의 단일 egocentric view 이미지를 stitching한 뒤 생성된 시야각 넓은 egocentric view를 기준으로 라이더 정보와 RGB 2D 이미지 정보를 합성하는 방법을 제안한다. 예를 들어, Front left image, Front image, Front right image를 image stitching(을 통해 하나의 wide RGB 2D 이미지(와이드 뷰 이미지)를 생성한다. 생성된 이미지를 ImageNet에 통과시켜 feature를 추출한 뒤 이를 range view 이미지에 warping한다. 그리고, 라이더 정보는 range view 이미지로 구성한 뒤 LidarRVNet에 통과시킨 후 앞서 생성한 range view 이미지와 range view fusion을 수행한다. 이를 추후 경로 예측에 사용한다.

[0055] 도 12는 앞에서 설명한 입력 요소 및 네트워크 구성 예시를 보인 것이다. 본 구조에서는 2개 이상의 단일 뷰 영상을 활용하여 멀티 뷰 영상을 생성하고 이를 활용하여 객체의 미래예측을 수행한다. 도 12와 같이 front left image, front image, front right image를 사용하여 시야각이 넓어진 영상과 라이더 정보를 합성하여 멀티 뷰를 사용할 수 있다. 이 외에도 다양한 뷰를 사용하여 다른 시야각을 합성할 수도 있다. 시야각이 넓어진 영상을 합성하고 그 후에는 기존의 미래 객체 예측 방법을 사용하여 미래 예측을 수행한다.

[0056] 이와 같이 본 발명에서는 2D 영상이 한정적인 시야각을 갖고 있는 문제를 해결하고자 2D 영상 외에도 라이더 정보를 활용한다. 따라서 영상과 라이더 정보로 이루어진 멀티 모달을 이루게 되고, 이를 통한 미래 객체의 예측을 수행하게 된다. 먼저 도 3 내지 도 5와 같이 라이더 정보를 처리하는 방법을 택할 수 있다. 라이더 정보를 처리하여 얻게 되는 출력 값과 RGB 영상을 활용하여 얻게 되는 출력 값 간의 융합을 하여 최종의 결과값을 얻게 된다. 이때 도 12와 같은 멀티 뷰 RGB 네트워크를 사용할 수도, 단일 RGB 영상을 사용하는 네트워크를 사용할 수도 있다.

[0057] 도 13은 pointpillar를 활용한 pillar feature net를 통해 라이더 정보를 처리하는 네트워크의 예시이다. Pointpillar(포인트필라)을 통해 (x,y,z)로 이루어져 있는 point cloud(포인트 클라우드)를 C 채널, W 넓이, H 높이를 가진 pseudo image(슈도 이미지)를 생성하고 컨볼루션 레이어를 통해 최종 벡터를 생성한다. 도 13에서는 단일 RGB영상을 처리하는 네트워크를 사용하는 예시를 보이고 RGB영상을 사용하여 마지막으로 출력되는 벡터에 라이더 모달의 정보를 쌓아서 fully connected 레이어를 통과하여 최종 미래 위치의 예측을 하게 된다. 혹은 기존 방법과 같이 각각의 모달을 융합할 시 최종 예측 결과 확률을 단순하게 더하는 구조를 사용할 수 있다.

[0058] 다음, 도 14의 멀티모달의 결과를 확률 분포로 융합하는 방법을 참조하면, RGB 모달과 라이더 모달의 각각 출력된 값을 융합 시, 융합 모델(mixture model)을 이용하여(S70) 가우시안 혼합 확률 분포를 생성한다(S80). 각각의 모달에서 출력된 N개의 미래 위치의 hypotheses를 2개의 fully-connected layer(fc)를 이용하여 soft assignment를 생성 및 추정한다. 추정된 soft assignments로 k개의 가우시안 확률 분포 모델의 파라미터 (π_k, μ_k, σ_k)를 계산하게 된다. 최종적으로 계산된 평균, 분산, 분포의 가중치를 이용하여 가우시안 혼합 확률 분포를 생성한다.

[0059] 부연하면, 객체의 미래 위치를 추정하기 위해 기존의 연구에서는 RGB 2D 이미지를 많이 활용해왔다. 본 발명에서는 여러 개의 단일 뷰의 영상을 융합하여 멀티 뷰 상황에서의 미래 객체 예측 방법을 설명한다. 배경설명의 종래 기술과 같이 미래 위치를 추정하기 위해서 과거의 3장의 이미지를 사용한다. 이 때 RGB 영상 외에도 다음과 같은 요소 중 하나 또는 그 이상의 조합을 입력으로 사용한다.

[0060] 1) 과거 프레임의 semantic segmentation (의미론적 분할) 영상 정보를 이용한다. 각 객체의 객체 정보 및 속성에 따라 장면에 나타나는 확률 및 움직임 속도가 서로 다르기 때문에 그러한 정보를 반영하기 위한 정보이다.

[0061] (2) 이와 유사하게, 과거 물체의 bounding box 위치에 물체의 클래스를 명시한 mask 이미지를 생성하여 객체의 정보 및 속성을 표현하는 정보를 입력으로 사용한다.

[0062] (3) RGB 2D 이미지와 라이더 정보를 합성한 range view 이미지를 이용한다. 또한 앞서 설명한 semantic segmentation 및 mask 이미지를 각 과거 몇 장씩 concatenate 하여 입력으로 사용한다.

[0063] (4) 자동차의 계획된 ego-motion을 입력으로 사용한다.

[0064] (5) 라이다 정보 또는 pointpillar와 같은 3차원 라이다 특성 정보를 입력으로 사용한다.

[0065] (6) 차선 등 도로 정보를 이용한다.

[0066] (1)에서 (6)을 입력으로 사용하는 네트워크는 fully-connected layer를 통해 M(=20)개의 bounding box hypotheses를 생성한다. 이러한 네트워크의 학습은 아래 수식과 같은 EWTA loss를 이용해 네트워크를 학습시킨다. EWTA loss는 다음 식과 같이 표현할 수 있다.

[0067] $$L_{FLN}=\sum_{k=1}^{K} w_k l(h_k,\hat{y}), l=L2\ norm$$

[0068] y는 ground-truth, h는 추정된 hypothesis이다. Ground-truth와 각 hypothesis의 L2 norm을 계산하게 된다. 여기서 w는 weight를 의미하며, 0 혹은 1이다.

[0069] K winners(best hypotheses)를 업데이트 시키며, K개의 weight는 1, M-K개의 weight는 0이 된다. 여기서, K를 M부터 1까지 계속해서 반으로 줄여 나가며 네트워크를 학습시켜 다양한 hypotheses가 출력될 수 있게 한다.

[0070] 한편, 이러한 멀티 모달 융합시 어텐션 구조를 활용할 수 있다. 도 16은 셀프 어텐션 개념을 도식화한 것이며, 도 17은 실제 행렬 기반의 셀프어텐션 연산 과정을 나타낸 것이다.

[0071] 딥 러닝(Deep learning) 기술에서 어텐션 기술(Attention Mechanism) 이란 현재 모델이 학습하고 있는 feature가 자기 자신 또는 다른 feature에 얼마만큼의 연관성이 있는지 스스로 계산하여, 그 중에서 유용한 feature를 강화하도록 학습하는 기술을 의미한다. 처음 자연어 처리에서 시작하여, 영상 처리에 관한 딥 러닝 모델에 활용이 되고 있다.

[0072] 아래와 같이 자연어 처리에서 개발한 셀프 어텐션(Self-attention) 기술에 관하여 설명한다. 아래에서 주요 예제는 자연어 처리를 이용해 설명하지만, 영상 및 비디오 등의 다양한 센서 데이터에 적용이 가능하다.

[0073] 인코더에 초기 입력된 단어의 각 단어 벡터들에 가중치 행렬을 곱해 Query벡터, Key벡터, Value벡터를 얻는다.

[0074] 기존의 어텐션 메커니즘과 동일하게, 주어진 Query에 대해 모든 Key와의 어텐션 스코어를 각각 구하고, 이후 해당 유사도를 가중치로 하여 Key와 매핑되어 있는 각각의 Value에 반영한다. 마지막으로 유사도가 반영된 Value를 모두 가중합하여 어텐션 값을 구한다.

[0075] 예를 들어서 I am a student라는 입력에서 student라는 단어의 셀프 어텐션을 계산하는 방법을 도면에 나타내었다. Student의 단어 벡터에 가중치 행렬 W^Q, W^K, W^V를 곱해서 각각 단어 student에 대한 Query벡터, Key벡터, Value벡터 $Q_{student}, K_{student}, V_{student}$를 얻는다. 그리고 나서, 두 백터의 내적을 이용하는 스케일드 닷-프로덕트 어텐션(Scaled dot-product Attention)을 수행해서 $scorefunction(q,k)=q \cdot k/\sqrt{d_k}$수식을 통해 어텐션 스코어를 얻는다. $Q_{student}$과 모든 Key벡터의 어텐션 스코어를 구한 후 어텐션 분포를 구한 후 해당 유사도를 가중치로 하여 Key와 같은 단어를 매핑하는 각각의 Value에 반영하고, 유사도가 반영된 Value벡터의 가중함으로 최종 어탠션 값(attention value)를 구한다.

[0076] 본 발명에서 어텐션을 활용하여 여러 모달이 효과적으로 융합하는 방법을 설명하면, 기존 멀티모달의 융합 방법에서는 배경기술의 설명과 같이 여러 모달을 채널별로 쌓거나 차원을 맞춘 벡터를 더하여 처리한다. 본 방법에서는 2D 이미지와 라이더 정보를 융합하기 위해 어텐션 구조를 활용하여 모달을 융합한다.

[0077] 시간 t에서 2D RGB 정보를 활용하여 출력한 벡터의 집합을 X_t^R이라 정의하고, 라이더 정보로부터 생성되는 벡터의 집합을 X_t^L이라 정의한다. $x^R \in X_t^R$, $x^L \in X_t^L$ 모두 동일한 크기의 차원을 갖고 있다. Query, Key, Value를 각각 다음 방법을 선택해 설정할 수 있다.

[0078] Query를 $x^R \in X_t^R$으로 설정하고, Key와 Value를 $x^L \in X_t^L$으로 설정,

[0079] Query를 $x^L \in X_t^L$으로 설정하고, Key와 Value를 $x^R \in X_t^R$으로 설정.

[0080] Ω_q에서의 K번째 Key에 대해서 $A(\bullet,\bullet)$로 표현되는 attention weight를 통해 출력 벡터는 다음 식과 같이

표현할 수 있다.

[0081] $$y_q = \sum_{k \in \Omega_q} A(\Phi_Q(x_q), \Phi_K(x_k)) \cdot \Phi_V(x_k)$$

[0082] Φ_Q, Φ_K, Φ_V는 Query, Key, Value의 값을 처리하는 선형 레이어이다. Attention weight를 통해 Query-Key값을 계산해서 가장 유사도가 높은 값으로 가중치 합을 구하게 된다. 위 식을 통해 얻은 출력 값을 활용하여 최종 예측을 하게 된다.

[0083] 도 15는 이러한 어텐션 구조를 멀티모달에 활용한 것이다. 도 15는 딥 러닝-어텐션 구조(A)를 이용하여 LiDAR 모달에서 출력으로 나온 벡터(L)와 RGB 모달에서 나오는 벡터(R) 간의 유사도를 통해 융합하는 네트워크 예시를 보이고 있다.

[0084] 이상과 같은 본 발명은 예시된 도면을 참조하여 설명되었지만, 기재된 실시 예에 한정되는 것이 아니고, 본 발명의 사상 및 범위를 벗어나지 않고 다양하게 수정 및 변형될 수 있음은 이 기술의 분야에서 통상의 지식을 가진 자에게 자명하다. 따라서 그러한 수정 예 또는 변형 예들은 본 발명의 특허청구범위에 속한다 하여야 할 것이며, 본 발명의 권리범위는 첨부된 특허청구범위에 기초하여 해석되어야 할 것이다.

부호의 설명

[0085] S10 : 객체 위치 예측

S20 : 가설 도출

S30 : 객체 위치 분포 예측

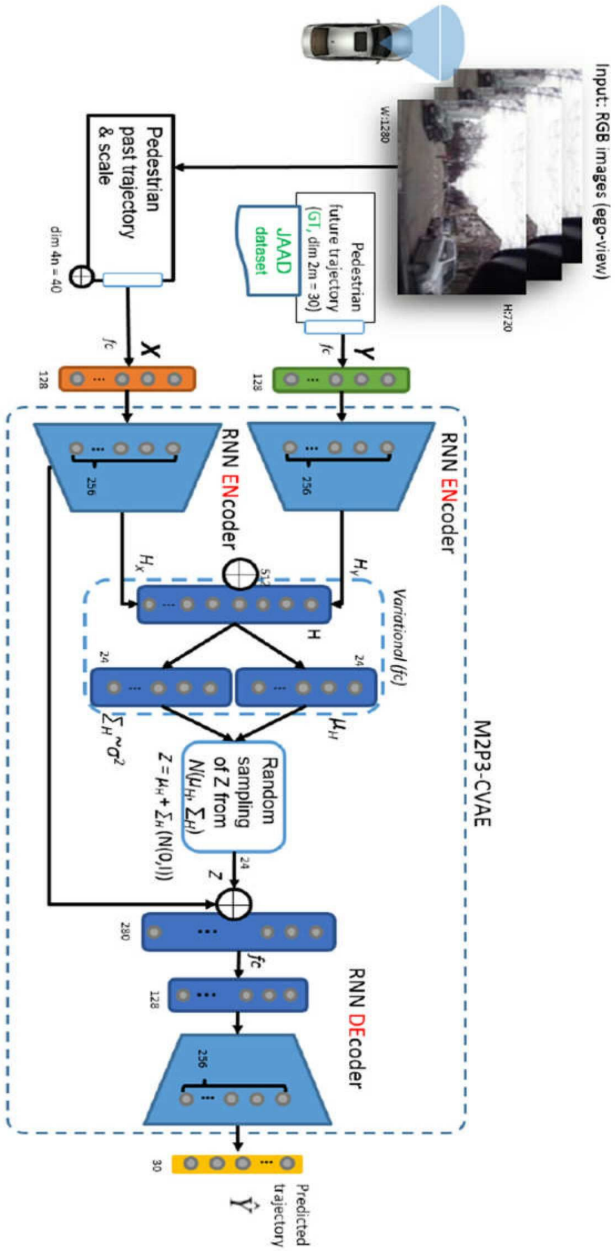
S40 : fitting

S50 : 융합 모델 생성

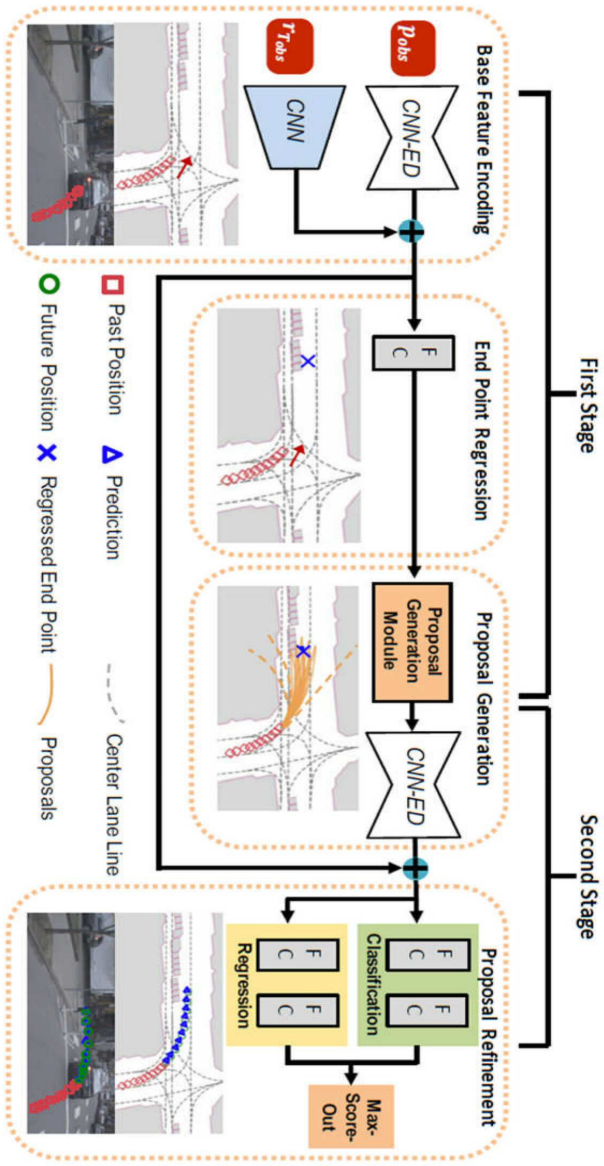
S60 : 가우시안 혼합 확률 분포 생성

도면

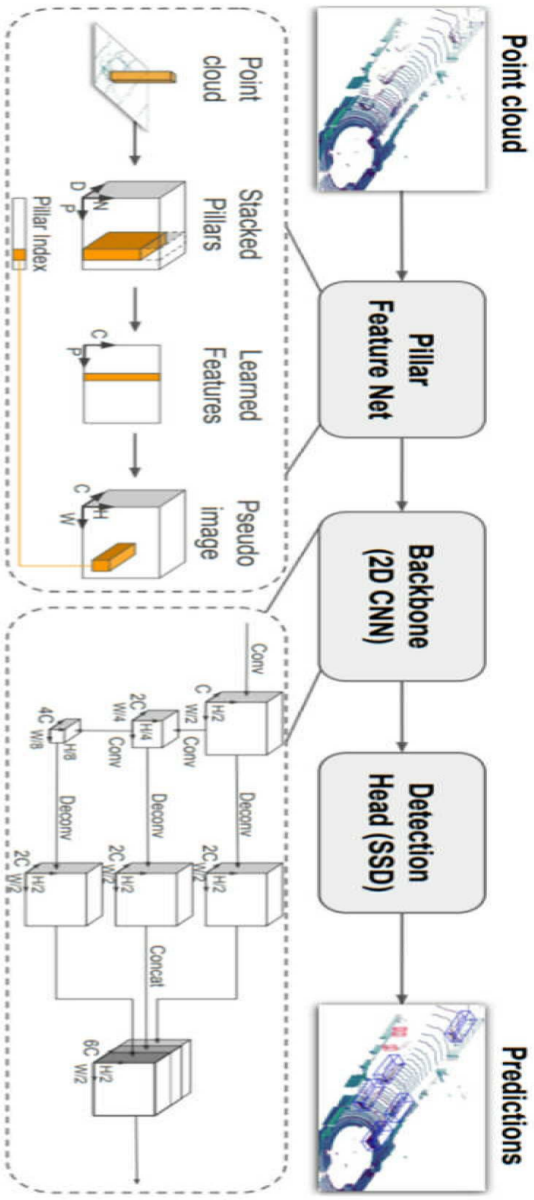
도면1


Input: RGB images (ego-view)
W:1280
H:720
Pedestrian past trajectory & scale
dim 4n = 40
Pedestrian future trajectory (GT, dim 2m = 30)
JAAD dataset
fc
X
Y
128
128
RNN ENcoder
RNN ENcoder
256
256
H_x
H_y
512
H
Variational (fc)
24
24
μ_H
$\sum_H \sim \sigma^2$
Random sampling of Z from $N(\mu_H, \sum_H)$
$Z = \mu_H + \sum_H (N(0,I))$
Z
24
280
fc
128
256
RNN DEcoder
M2P3-CVAE
30
$\hat{Y}$
Predicted trajectory

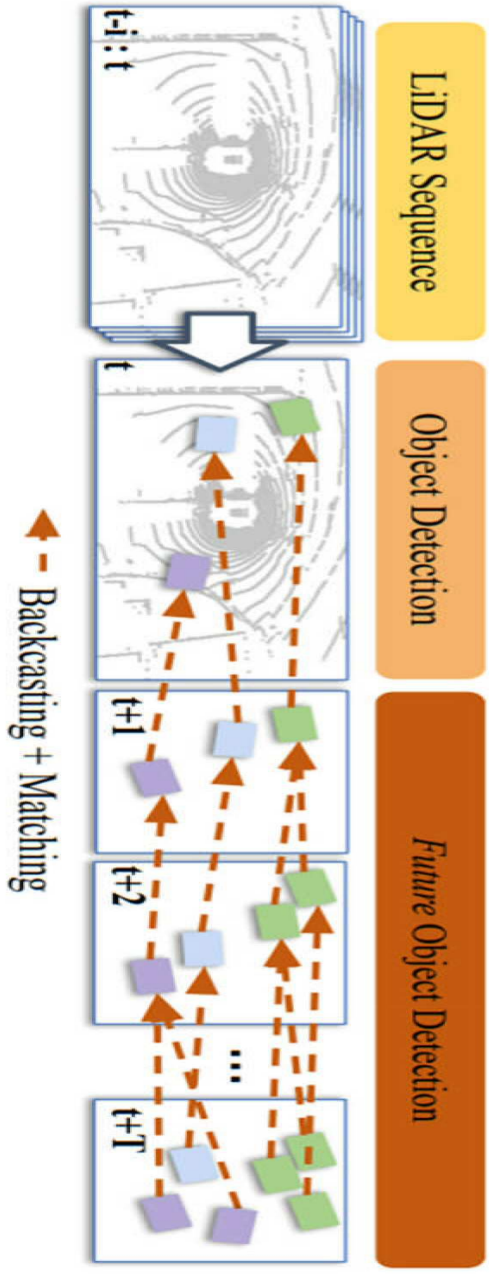
도면2


Base Feature Encoding
CNN-ED
CNN
End Point Regression
FC
Proposal Generation
Proposal Generation Module
CNN-ED
Proposal Refinement
FC
FC
Classification
FC
FC
Regression
Max-Score-Out
First Stage
Second Stage
Past Position
Prediction
Center Lane Line
Future Position
Regressed End Point
Proposals

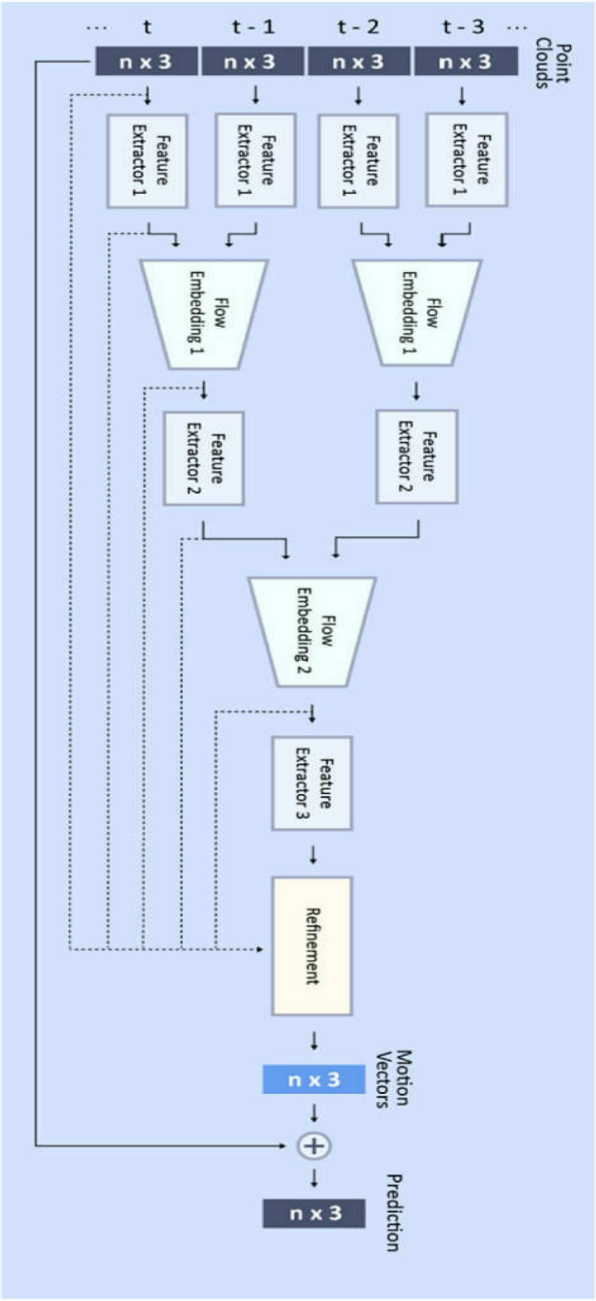
도면3


Point cloud
Pillar Feature Net
Backbone (2D CNN)
Detection Head (SSD)
Predictions
Point cloud
Stacked Pillars
Pillar Index
Learned Features
Pseudo image
Conv
Deconv
Concat

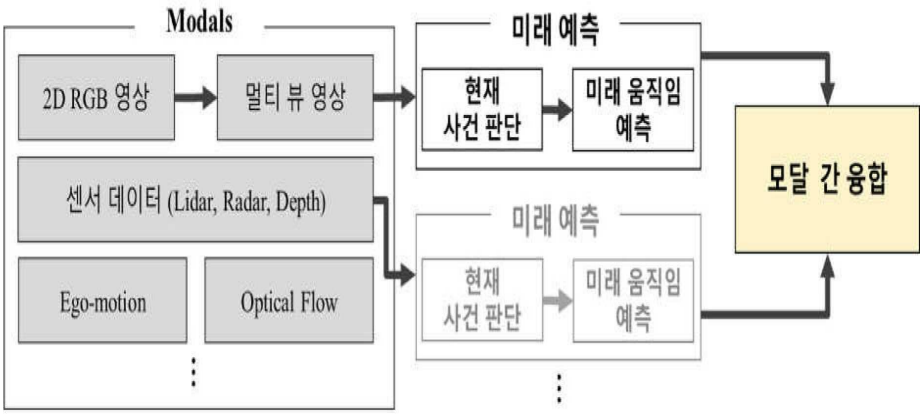
도면4


LiDAR Sequence
Object Detection
Future Object Detection
t-i : t
t
t+1
t+2
...
t+T
Backcasting + Matching

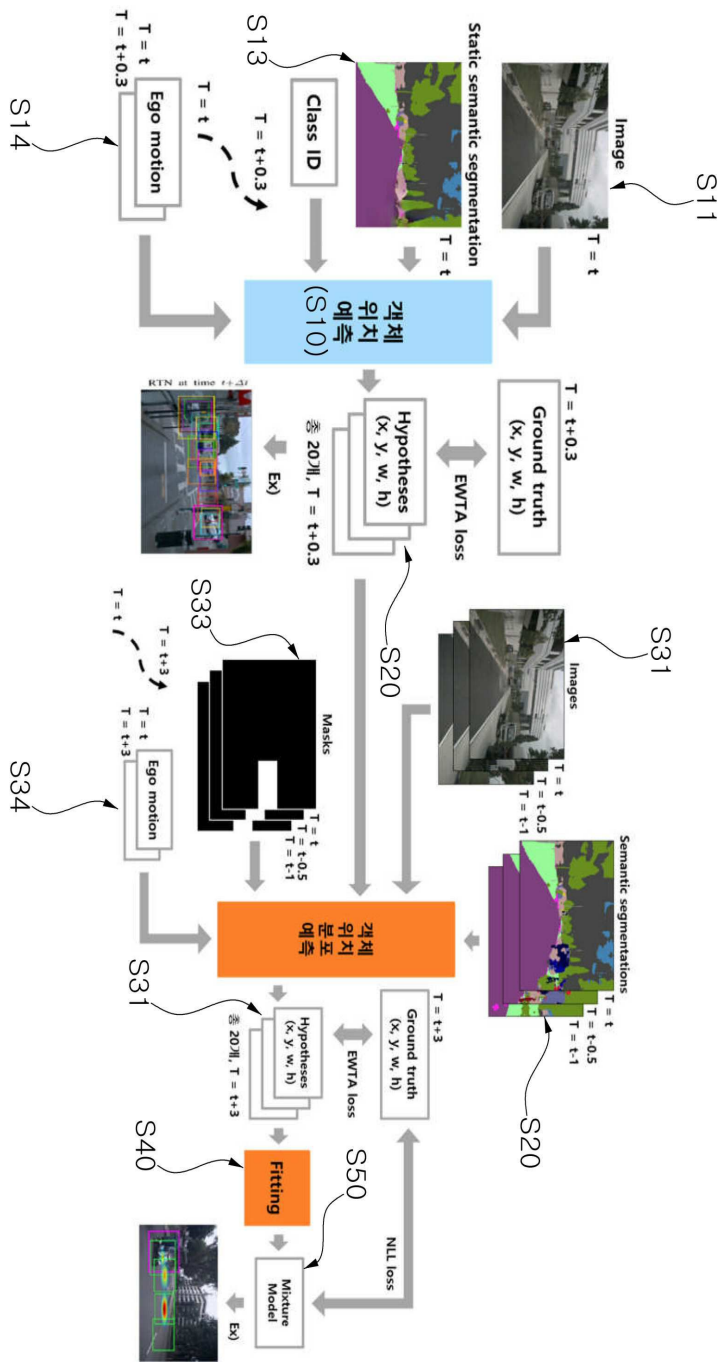
도면5


t
t - 1
t - 2
t - 3
n x 3
Point Clouds
Feature Extractor 1
Flow Embedding 1
Feature Extractor 2
Flow Embedding 2
Feature Extractor 3
Refinement
Motion Vectors
Prediction

도면6


Modals
2D RGB 영상
멀티 뷰 영상
센서 데이터 (Lidar, Radar, Depth)
Ego-motion
Optical Flow
미래 예측
현재 사건 판단
미래 움직임 예측
미래 예측
현재 사건 판단
미래 움직임 예측
모달 간 융합

도면7


S11
Image
T = t
S13
Static semantic segmentation
T = t
Class ID
T = t+0.3
T = t
S14
Ego motion
T = t
T = t+0.3
객체 위치 예측 (S10)
Hypotheses (x, y, w, h)
총 20개, T = t+0.3
Ex)
EWTA loss
Ground truth (x, y, w, h)
T = t+0.3
S20
S31
Images
T = t
T = t-0.5
T = t-1
Semantic segmentations
T = t
T = t-0.5
T = t-1
S20
S33
Masks
T = t
T = t-0.5
T = t-1
T = t
T = t+3
S34
Ego motion
T = t
T = t+3
객체 위치 분포 예측
S31
Hypotheses (x, y, w, h)
총 20개, T = t+3
EWTA loss
Ground truth (x, y, w, h)
T = t+3
S40
Fitting
S50
NLL loss
Mixture Model
Ex)

도면8

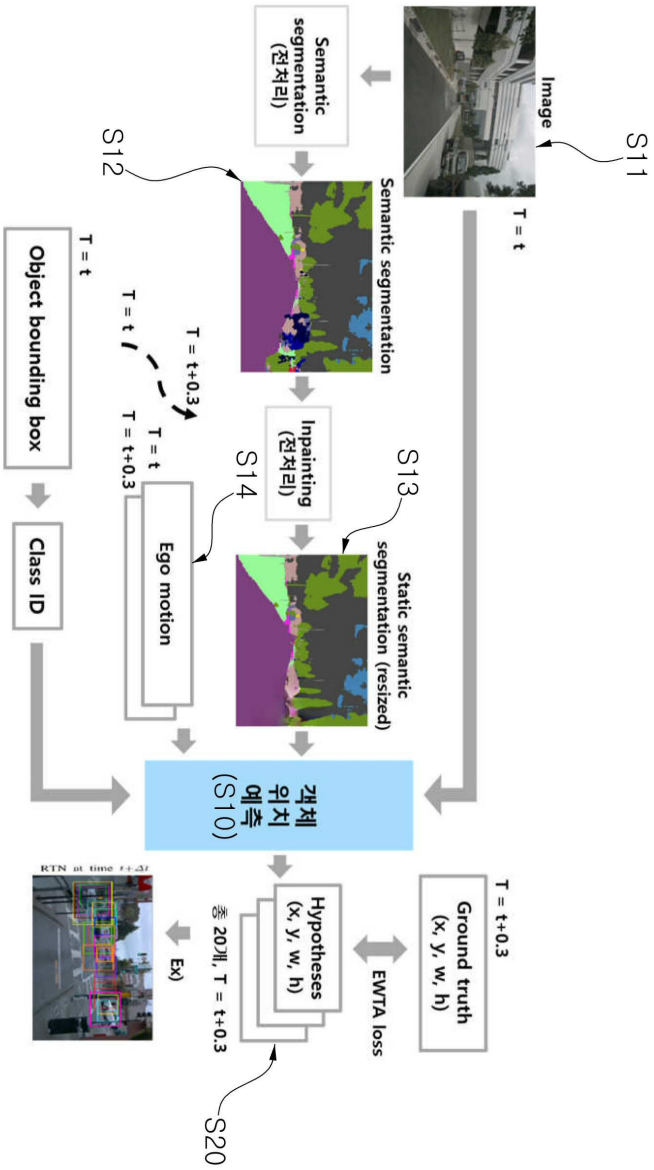
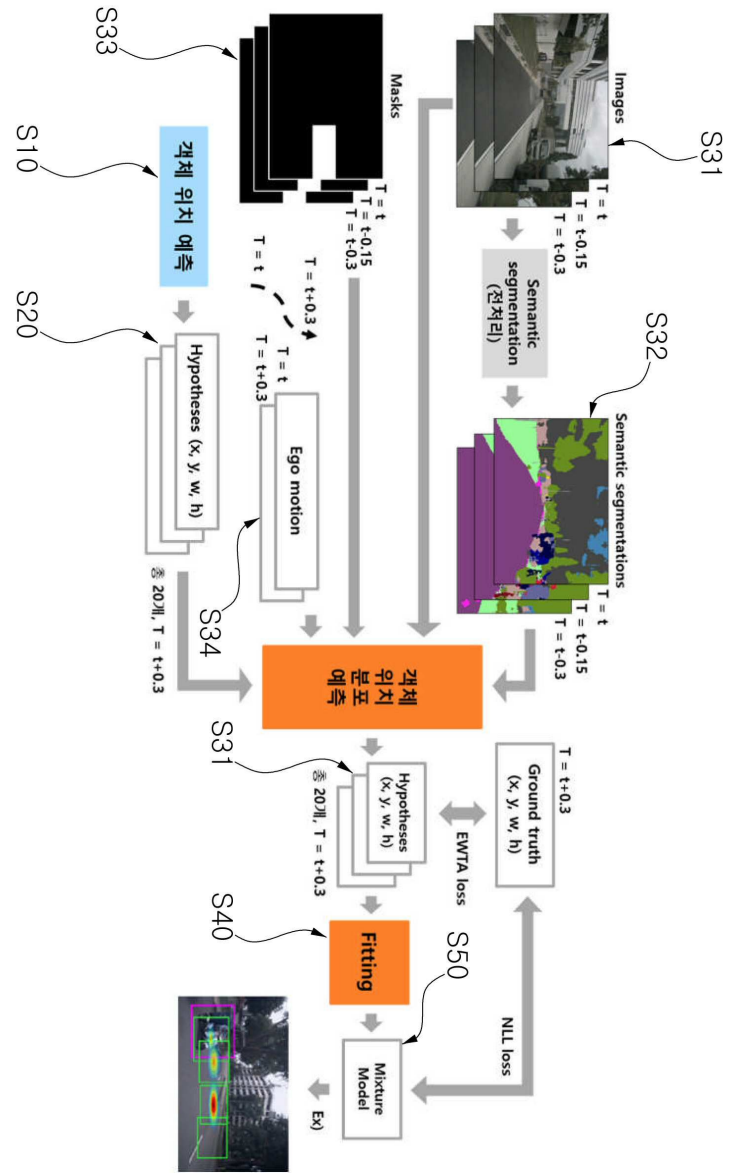
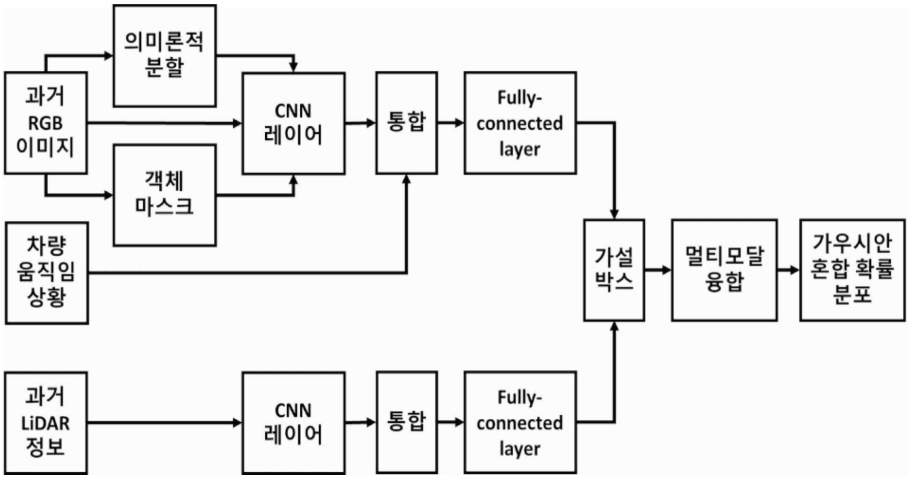


Image
S11
T = t
Semantic segmentation (전처리)
S12
Semantic segmentation
Inpainting (전처리)
S13
Static semantic segmentation (resized)
S14
Ego motion
T = t
T = t+0.3
T = t
T = t+0.3
T = t
Object bounding box
Class ID
객체 위치 예측 (S10)
T = t+0.3
Ground truth (x, y, w, h)
EWTA loss
Hypotheses (x, y, w, h)
총 20개, T = t+0.3
S20
Ex)

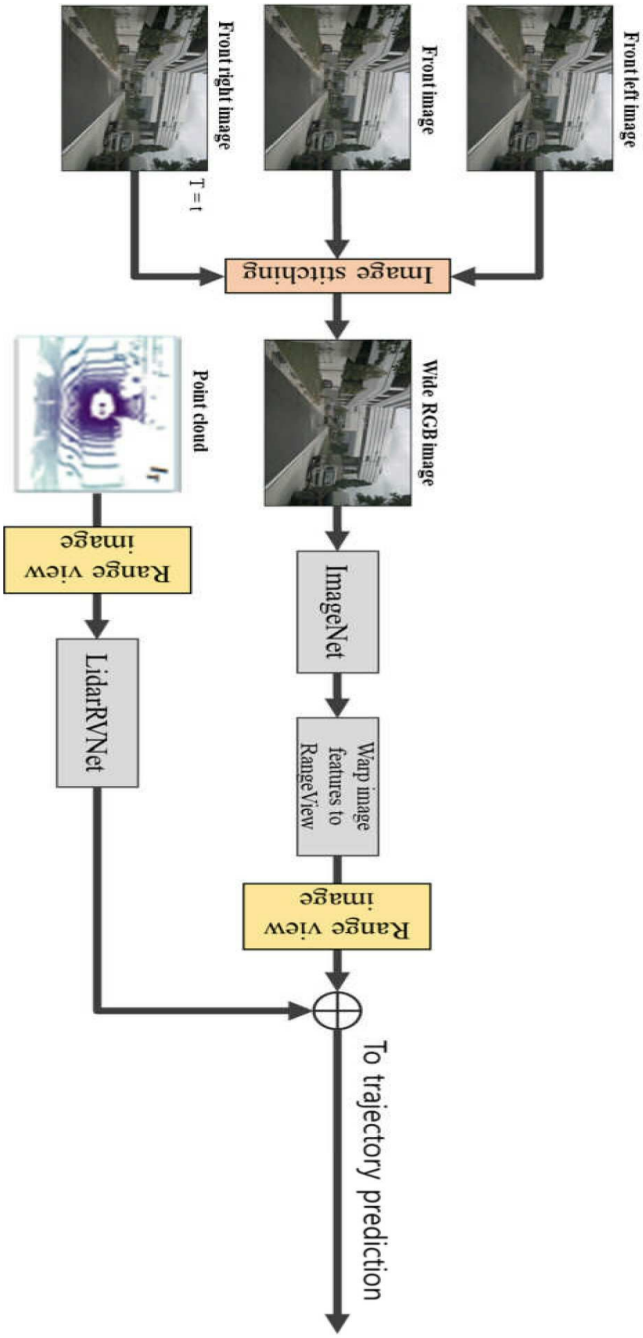
도면9


S33
Masks
T = t
T = t-0.15
T = t-0.3
S10
객체 위치 예측
T = t
T = t+0.3
T = t
T = t+0.3
S20
Hypotheses (x, y, w, h)
총 20개, T = t+0.3
Ego motion
S34
Images
S31
T = t
T = t-0.15
T = t-0.3
Semantic segmentation (전처리)
S32
Semantic segmentations
T = t
T = t-0.15
T = t-0.3
객체 위치 분포 예측
S31
Hypotheses (x, y, w, h)
총 20개, T = t+0.3
EWTA loss
Ground truth (x, y, w, h)
T = t+0.3
S40
Fitting
S50
NLL loss
Mixture Model
Ex)

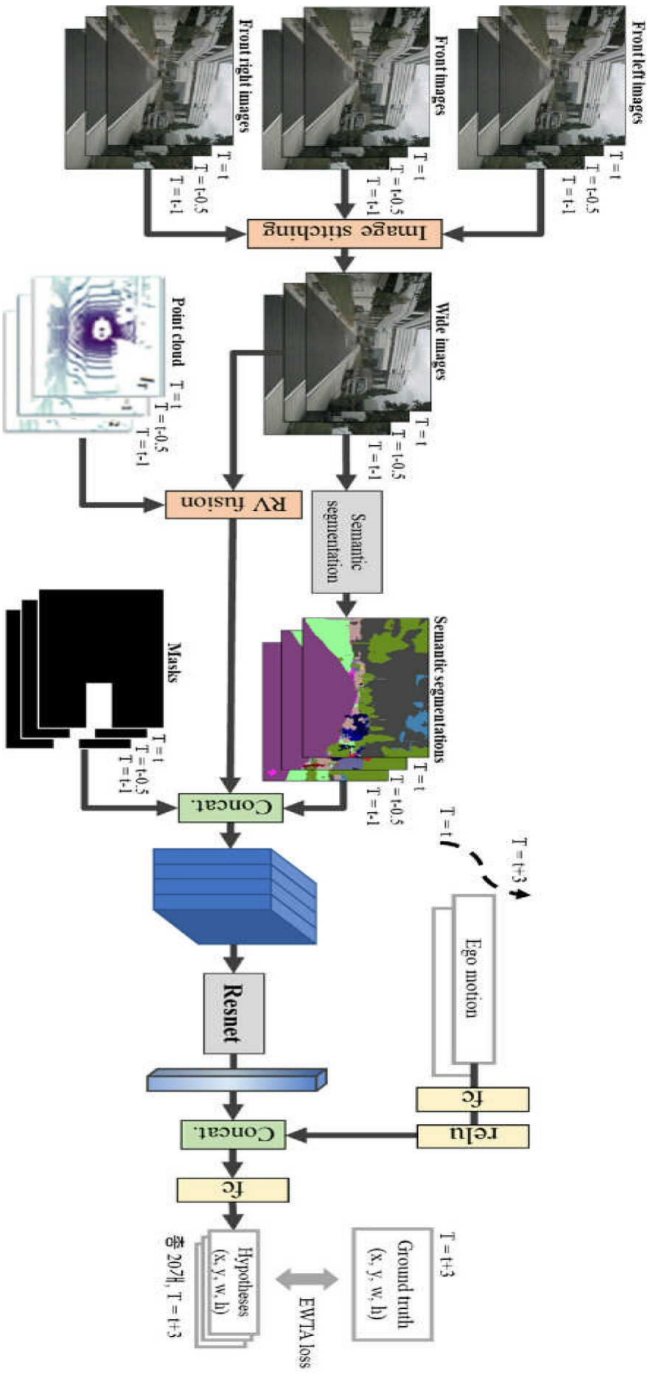
도면10


의미론적 분할
과거 RGB 이미지
CNN 레이어
통합
Fully-connected layer
객체 마스크
차량 움직임 상황
가설 박스
멀티모달 융합
가우시안 혼합 확률 분포
과거 LiDAR 정보
CNN 레이어
통합
Fully-connected layer

도면11


Front left image
Front image
Front right image
T = t
Image stitching
Wide RGB image
Point cloud
ImageNet
Range view image
LidarRVNet
Warp image features to RangeView
Range view image
To trajectory prediction

도면12


Front left images
Front images
Front right images
T = t
T = t-0.5
T = t-1
Image stitching
Wide images
Point cloud
RV fusion
Semantic segmentation
Semantic segmentations
Masks
Concat.
Resnet
Concat.
fc
relu
fc
Ego motion
T = t+3
Ground truth
(x, y, w, h)
EWTA loss
Hypotheses
(x, y, w, h)
총 20개, T = t+3

도면13

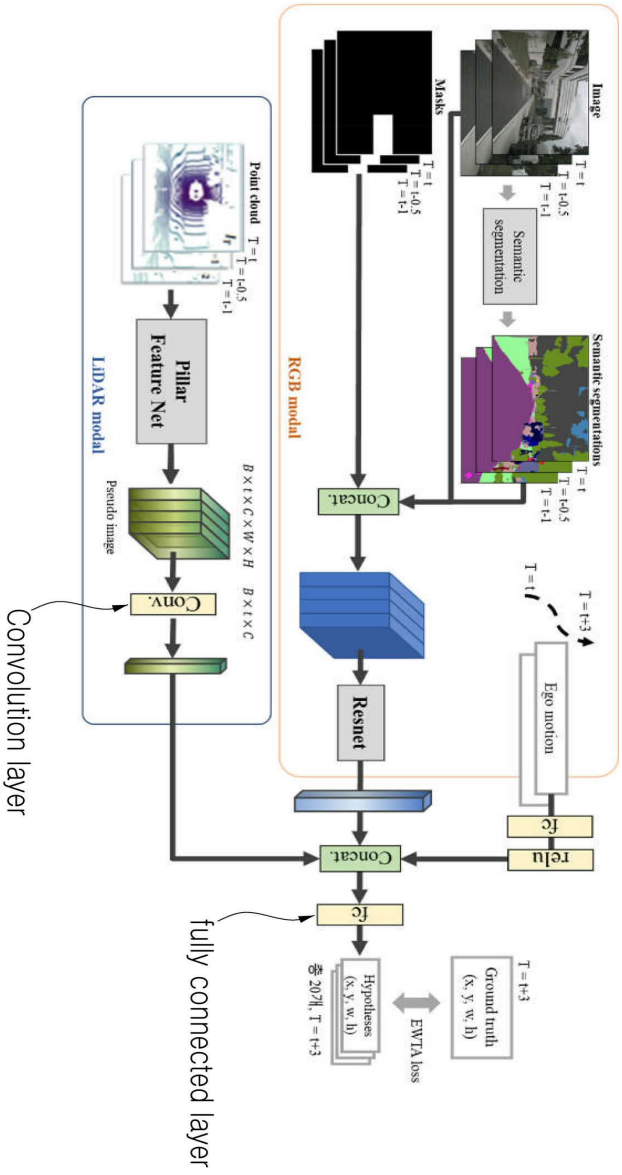
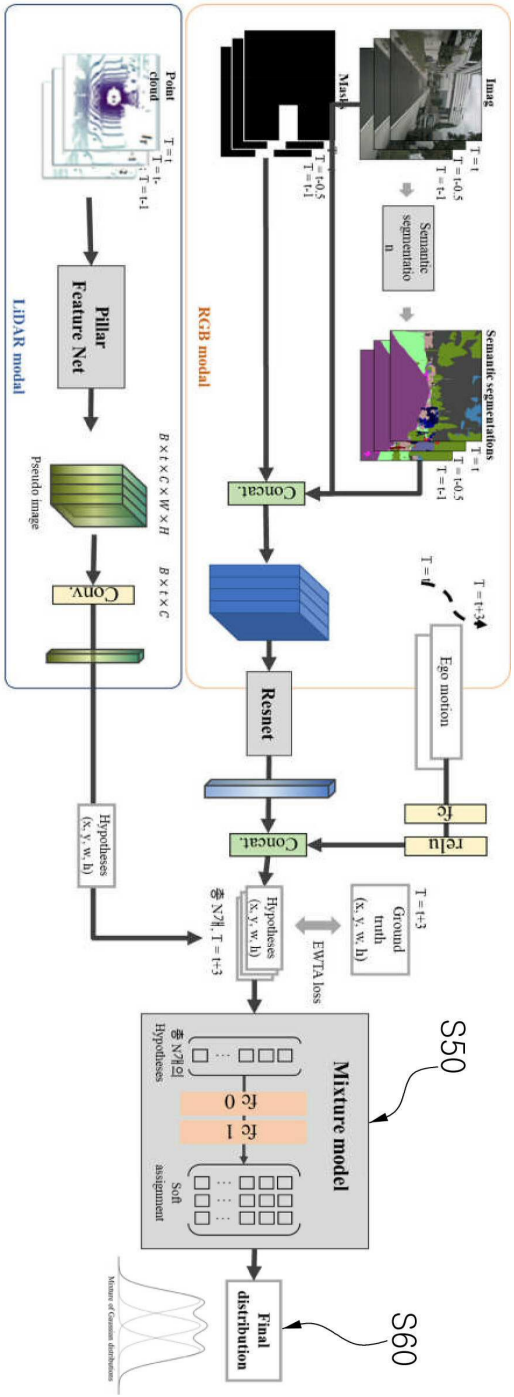


Image
Masks
Semantic segmentation
Semantic segmentations
RGB modal
Concat.
Resnet
Ego motion
fc
relu
Concat.
fc
Hypotheses (x, y, w, h)
Ground truth (x, y, w, h)
EWTA loss
Point cloud
Pillar Feature Net
LiDAR modal
Pseudo image
Conv.
Convolution layer
fully connected layer

도면14


Point cloud
Pillar Feature Net
LiDAR modal
Pseudo image
Conv.
Hypotheses (x, y, w, h)
RGB modal
Mask
Image
Semantic segmentation
Semantic segmentations
Concat.
Resnet
Concat.
Ego motion
fc
relu
Hypotheses (x, y, w, h)
Ground truth (x, y, w, h)
EWTA loss
Mixture model
Hypotheses
fc 0
fc 1
Soft assignment
Final distribution
S50
S60

도면15

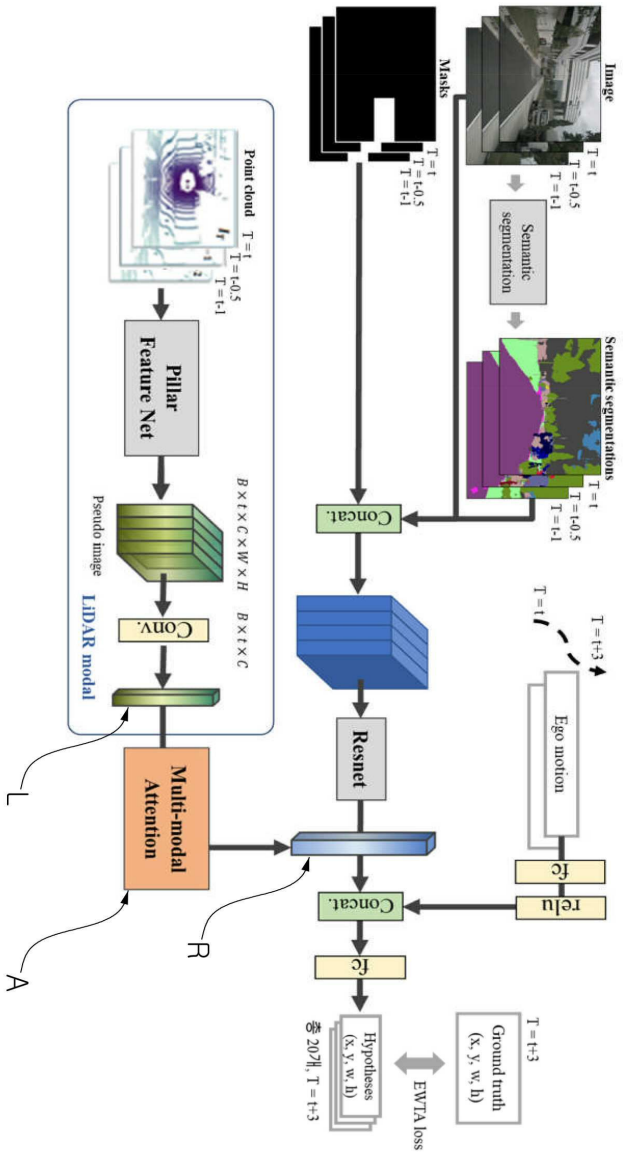
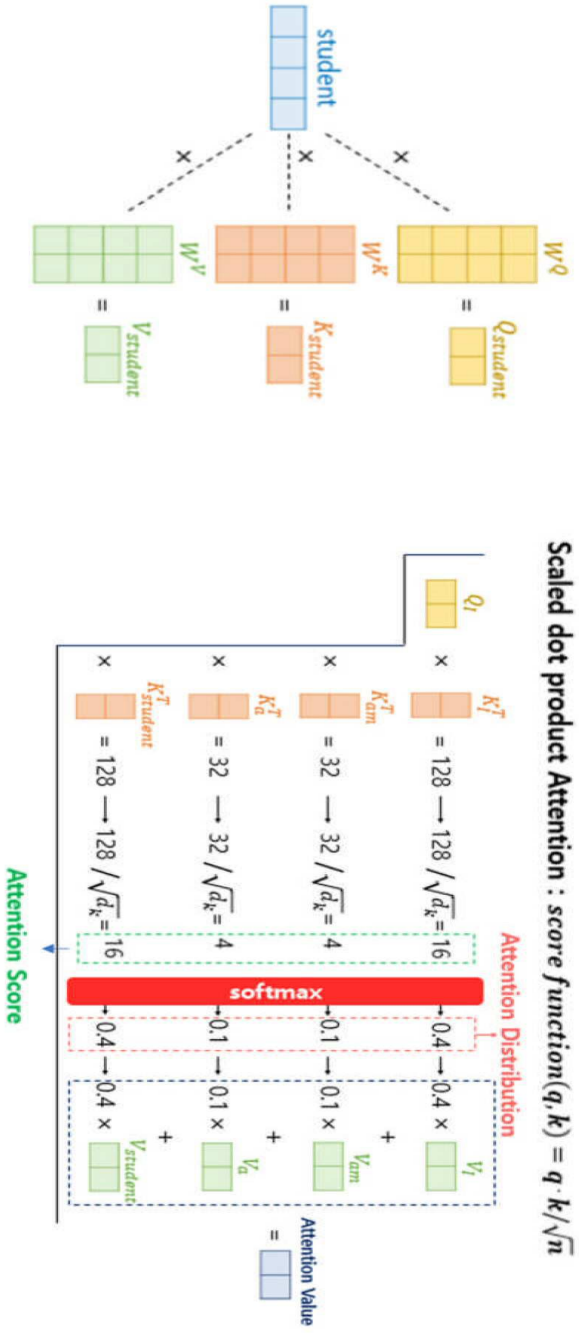
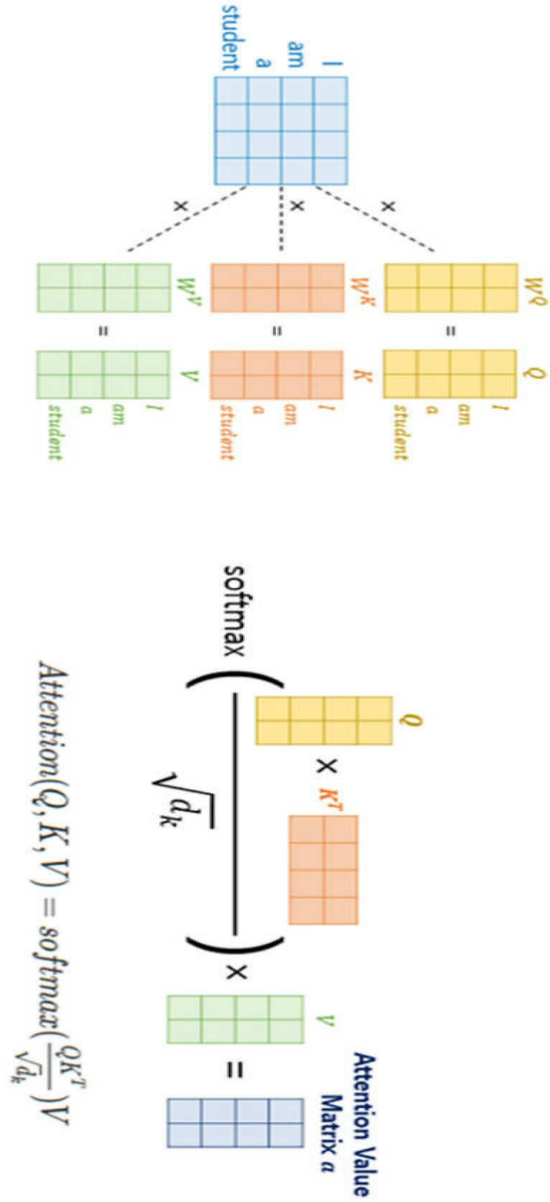


Image
T = t
T = t-0.5
T = t-1
Masks
Semantic segmentation
Semantic segmentations
Concat.
Point cloud
Pillar Feature Net
Pseudo image
Conv.
LiDAR modal
B × t × C × W × H
B × t × C
Resnet
Ego motion
T = t+3
fc
relu
Concat.
fc
Multi-modal Attention
L
A
R
Hypotheses (x, y, w, h)
예: 20개, T = t+3
EWTA loss
Ground truth (x, y, w, h)

도면16


student
W^V
W^K
W^Q
$V_{student}$
$K_{student}$
$Q_{student}$
Scaled dot product Attention : $score\ function(q,k) = q \cdot k/\sqrt{n}$
q_i
K_i^T
K_{am}^T
K_a^T
$K_{student}^T$
= 128 ⟶ 128 /$\sqrt{d_k}$= 16
= 32 ⟶ 32 /$\sqrt{d_k}$= 4
= 32 ⟶ 32 /$\sqrt{d_k}$= 4
= 128 ⟶ 128 /$\sqrt{d_k}$= 16
Attention Distribution
softmax
0.4
0.1
0.1
0.4
V_i
V_{am}
V_a
$V_{student}$
Attention Score
Attention Value

도면17


I
am
a
student
W^Q
W^K
W^V
Q
K
V
Attention Value Matrix a
softmax
$\sqrt{d_k}$
K^T
$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$