논문 22-47-02-12 The Journal of Korean Institute of Communications and Information Sciences '22-02 Vol.47 No.02
https://doi.org/10.7840/kics.2022.47.2.300

레이더 신호와 머신러닝을 이용한 비가시 공간에서의 삼차원 객체 인식 연구

김 건 우•, 이 상 원*, 손 하 영*, 최 계 원°

A Study on 3D Object Detection in Invisible Area Using Radar Signal and Machine Learning

Gon-Woo Kim•, Sang-Won Lee*, Ha-Young Son*, Kae-Won Choi°

요 약

비가시 공간의 물체를 탐지하는 기술은 군사적, 인명구조, 자율 주행 등의 다양한 목적에서 주목받고 있다. RF 레이더 신호는 벽을 투과하여 물체를 측정할 수 있으므로 이 일을 수행하기 적합한 센서 형태로 꼽히고 있다. 본 논문에서는 다중 송·수신 안테나와 초광대역 레이더 칩을 통해 데이터 수집 실험 환경을 구성한다. 구성한 환경을 이용하여 수집한 신호를 데이터셋으로 사용하고 해당 데이터셋을 Transformer 모델의 입력으로 사용하여 비가시 공간의 물체를 Bird-Eye-View Bounding Box를 통한 3D Object Detection을 수행하여 물체의 위치 추정을 위한 알고리즘을 제시한다.

Key Words : Radar, MIMO, Attention, Transformer, Object Detection

ABSTRACT

Technology for detecting objects in invisible spaces is attracting attention for various purposes such as military, lifesaving, and autonomous driving. RF radar signals are considered as suitable sensor types for performing this task because they can measure objects through walls. In this paper, a data collection experiment environment is constructed through a MIMO(Multi-In-Multi-Out) antenna and a Ultra-Wideband radar chip. Using signals collected using the configured environment as datasets and the corresponding dataset as input of the Transformer model, 3D object detection through Bird-Eye-View Bounding Box is performed to present algorithms for object position estimation.

Ⅰ. 서 론

현대에 들어 사람이 직접 확인하기 어려운 비가시 공간에서 물체를 탐지하는 기술의 중요성이 점점 커지고 있다. 먼저 군사적으로, 건물 구조와 군사 작전의 상황이 다양해지는 현대에서 엄폐 영역에서의 적군을 탐지하는 것은 작전의 성공에 큰 영향을 끼친다. 건물이 무너져 있거나 화학 화재와 같이 사람이 투입

※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2020-0-00973, VR·AR 콘텐츠 비가시 영역 영상 복원 기술 개발)을 받아 수행된 연구임.

• First Author : Sungkyunkwan University Department of Electrical and Computer Engineering, rjsdn4444@skku.edu, 학생회원
° Corresponding Author : Sungkyunkwan University Department of Electrical and Computer Engineering, kaewonchoi@skku.edu, 종신회원
* Sungkyunkwan University, Department of Electrical and Computer Engineering, 학생회원
논문번호 : 202112-327-D-RU, Received December 3, 2021; Revised December 21, 2021; Accepted December 26, 2021

되어 탐사하기 어려운 재난 상황에서도 비가시 영역 탐지 기술은 구조 대상자들의 위치를 빠르게 찾아내어 빠른 구조 활동을 가능하게 만든다. 또한, 자율 주행 분야에서도 비가시 영역 탐지의 역할은 중요하다. 자율 주행 차량에서의 주요 과제 중 하나는 주변의 물체를 감지하는 것으로, 건물과 차량 등 여러 장애물이 밀집해 있는 도시에서 비가시 탐지의 중요성이 특히 커지고 있다.

최근 비가시 영역에서 물체의 위치를 정확히 탐지하고 활용을 늘리기 위한 삼차원 물체 인식이 주목받고 있다. 특히 자율주행 분야에서 카메라와 라이다/레이더 센서를 활용한 삼차원 물체 인식 연구가 활발히 이루어지고 있다. 기존의 삼차원 물체 인식 알고리즘은 주로 라이다가 출력하는 삼차원 포인트 클라우드 형태를 기반으로 개발되었다[1],[2]. 하지만 삼차원 포인트 클라우드 형태의 삼차원 물체 인식 알고리즘은 장애물을 투과하여 비가시 영역의 물체를 인식하는 목적과 상이하다. 레이더 센서는 라이다 센서에 비해 비, 안개와 같은 날씨와 조명 조건에 영향을 받지 않는 장점을 가지고 있어[3] 최근 레이더 센서를 기반으로 적합한 네트워크에 대해 연구가 이루어지고 있다[3,4]. [3]에서는 레이더 신호를 range-azimuth heatmap으로 변환하여 학습 데이터로 활용한다. [4]에서는 레이더 신호를 포인트 클라우드로 변환한 데이터와 카메라를 통한 RGB 이미지를 학습 데이터로 활용한다. 하지만 본 논문에서는 트랜스포머[5]를 이용하여 사전 처리 및 센서 융합을 이용하지 않고 레이더 신호를 그대로 이용한 비가시 영역에서의 물체 인식 알고리즘을 제안한다.

RF(Radio Frequency) 신호는 주파수에 따라 투과도가 달라져 다양한 재질의 물체를 측정할 수 있는 장점이 있으며, 또한 주파수 조정을 통해 벽을 투과할 수 있도록 만들 수 있다. 이를 활용한 RF 신호는 비가시 영역에서의 물체를 탐지할 수 있는 특징을 가지고 있다[6]. 이러한 RF 신호의 특성을 통해 송신부의 안테나에서 송출한 신호가 시야를 방해하는 장애물을 통과하고, 장애물 뒤에 있는 물체에 RF 신호가 반사된다. 반사된 신호는 수신부의 안테나를 통해 데이터 형태로 신호를 저장할 수 있다. 이때, 공간적 정보를 더 많이 수집하기 위해서 송·수신부에 다수의 안테나를 배치할 수 있는데, 다중 입출력을 뜻하는 MIMO(Multiple Input Multiple Output, 다중 안테나 기술)을 통해 RF 신호의 해상도를 높이고 다양한 안테나 구성을 통해 물체에 대한 공간적 정보를 비교적 정확하게 추정할 수 있도록 할 수 있다[7].

본 논문에서는 UWB(Ultra-Wide Bandwidth) 레이더 칩과 다중 안테나 기술을 활용하여 레이더 신호를 통해 공간에 대한 더 많은 정보를 구성한 시스템을 통해 자체적으로 수집하고, 수집한 정보를 딥러닝에 적용하여 물체 및 사람을 인식하고 삼차원 위치 추정 알고리즘을 구현한다. 학습을 위한 모델은 어텐션 기법을 활용한 트랜스포머 모델을 기반으로 한다. 트랜스포머 모델을 통해 레이더 신호의 캘리브레이션 및 이미징 등의 사전 처리를 하지 않고 1차원의 샘플링된 신호를 트랜스포머의 어텐션을 사용하여 안테나 조합 및 신호 간의 관계성 학습을 통한 물체 인식을 수행하도록 한다. 훈련을 위한 Label은 OptiTrack 사의 모션 캡쳐 시스템을 이용해 삼차원 공간 좌표를 수집하고 Bird-Eye-View를 통해 삼차원 물체의 위치를 2차원의 Bounding Box 형태로 표현한다. 그리고 Hungarian Algorithm을 통해 Prediction Box와 Label Box를 매칭하여 학습에 활용한다.

Ⅱ. 관련 연구

2.1 레이더를 이용한 객체 인식

레이더 데이터를 이용한 객체 인식 연구[6,8-11]가 활발히 이루어지고 있다. [8], [9], [10]에서 FMCW (Frequency-Modulated Continuous Wave) 레이더를 사용하여 객체 인식 연구를 진행하였다. 하지만 본 논문에서 사용하는 UWB 레이더는 FMCW 레이더에 비해 넓은 대역폭을 사용하여 얻을 수 있는 정보의 양이 많다. 예를 들어 [8]의 FMCW 레이더의 대역폭은 150MHz인데 반해 본 논문의 UWB 레이더의 대역폭은 3GHz에 이른다. 본 논문에서는 이러한 UWB 레이더를 사용하여 시스템을 구성하고 구성한 시스템을 통해 자체 데이터를 수집한다. 또한 [8], [9], [10]에서는 레이더 신호를 그대로 사용하지 않고 이미지로 변환하여 학습 데이터로 활용한다. MIMO 레이더에서의 이미징 작업은 복잡한 사전 캘리브레이션 과정이 필요하며 얻어진 신호를 실시간으로 이미지로 변환하는 과정이 요구된다. 하지만 본 논문에서는 레이더 신호의 사전 처리 과정 없이 신호 그대로를 데이터로 사용하여 객체 인식을 진행한다. 이를 통해 레이더 신호에 대한 캘리브레이션 및 이미징 추가 작업 없이 레이더 신호 그대로 데이터로 사용하여 알고리즘의 활용 가능성을 높인다.

[6], [11]에서는 UWB 레이더를 활용하여 객체 인식을 수행한다. [6]은 비가시 영역에서의 객체 인식 연구이지만 [11]의 객체 인식은 가시 상황에서의 객체

인식 연구이다. 또한 [6], [11] 모두 딥러닝을 이용한 객체 인식이 아니므로 딥러닝과 UWB 레이더 데이터를 사용하여 비가시 영역에서의 객체 인식을 수행한 본 논문과 차이가 있다.

2.2 Seq2Seq(Sequence-to-Sequence)

Seq2Seq(Sequence-to-Sequence) 모델[12]은 입력된 시퀀스로부터 다른 도메인의 시퀀스를 출력하는 모델이다. 해당 모델은 인코더, 컨텍스트 벡터, 디코더의 구조를 가진다. 인코더는 입력 문장의 모든 단어를 차례대로 입력받은 뒤에 마지막에 입력받은 모든 단어 정보들을 대표하는 하나의 벡터를 출력으로 가진다. 이때 모든 단어 정보를 대표하는 하나의 벡터를 컨텍스트 벡터라 한다. 그리고 디코더는 해당 컨텍스트 벡터를 입력으로 받아 번역된 단어를 한 개씩 차례대로 출력하는 구조를 가진다.

인코더에서 입력 문장은 단어 토큰화를 통해 단어 단위로 쪼개지고 각 단어 토큰은 각각 RNN(Recurrent Neural Network) 셀의 입력이 된다. 인코더에서 [그림 1]과 같이 RNN 셀은 토큰화된 단어를 순차적으로 입력받은 뒤에 인코더 RNN 셀의 마지막 시점의 은닉 상태인 컨텍스트 벡터를 디코더의 RNN 셀의 초기 은닉 상태 입력값으로 사용한다. 디코더에서는 학습 과정과 테스트 과정으로 동작이 나뉘는데, 학습 과정에서의 디코더는 인코더에서 받은 컨텍스트 벡터를 이용하여 Teacher Forcing을 통해 실제 번역된 단어를 입력으로 받아 학습하게 된다. 그리고 테스트 과정에서는 입력받은 컨텍스트 벡터와 문장의 시작을 의미하는 <sos> 토큰을 이용하여 첫 번째 단어를 예측하고 예측한 단어를 다음 시점의 RNN 셀의 입력으로 받아 다음 단어를 예측하게 된다.

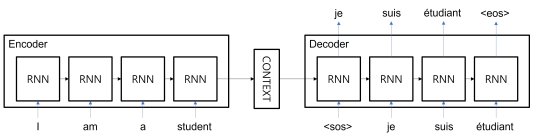


그림 1. Seq2Seq 모델
Fig. 1. Seq2Seq model

2.3 어텐션

상기 Seq2Seq 모델에는 크게 두 가지 문제가 존재한다. 첫 번째로 하나의 고정된 크기의 벡터에 모든 정보를 압축하므로 정보의 손실이 필연적으로 발생한다. 그리고 RNN을 사용함으로써 가지는 고질적인 문제인 Vanishing Gradient 문제가 있다. 따라서 입력 문장의 길이가 길어질수록 번역의 품질이 현저히 떨어지는 문제를 가지고 있다. 해당 문제를 해결하기 위해 어텐션[13]을 통해 입력 문장과 번역 문장의 단어 간의 관계성을 파악하여 문장의 길이가 길어져도 관계성 정보를 통해 정확도가 떨어지는 것을 막을 수 있다.

어텐션에서의 어텐션 함수는 [그림 2]와 같이 주어진 Query에 대해서 모든 Key와의 유사도를 각각 구한다. 구한 유사도를 키와 매핑되어 있는 각각의 Value에 반영한다. 그리고 유사도가 반영된 Value들은 모두 더해주고 해당 더해준 값을 어텐션 값이라 한다.

Seq2Seq 모델에서의 어텐션 기법은 [그림 3]과 같다. Seq2Seq 모델의 문제점을 해결하기 위해 인코더의 모든 입력 단어들의 정보를 출력 단어 예측을 위해 반영하는 구조를 가진다. 어텐션 값을 구하는 데에 먼저 어텐션 스코어를 구하는데, 어텐션 스코어는 현재 디코더의 시점이 t라 할 때 시점 t에서의 단어를 예측하기 위해 인코더의 모든 은닉 상태와 현재 시점의 디코더 은닉 상태 s_t의 유사도를 의미한다. [그림 3]에서 Softmax 층 이전에서 일반적으로 인코더의 은닉 상태와 디코더의 은닉 상태의 내적 연산 이후에 Softmax 층을 거친 값을 어텐션 스코어라 한다.

이렇게 구한 어텐션 스코어는 각 인코더의 은닉 상

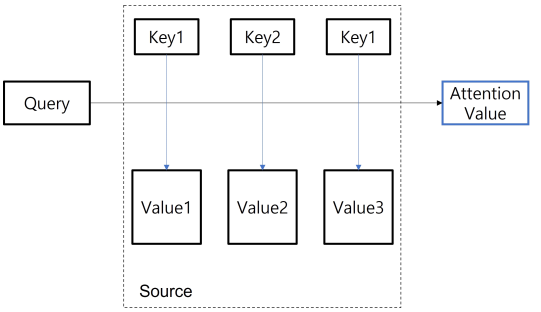


그림 2. 어텐션 함수
Fig. 2. Attention Function

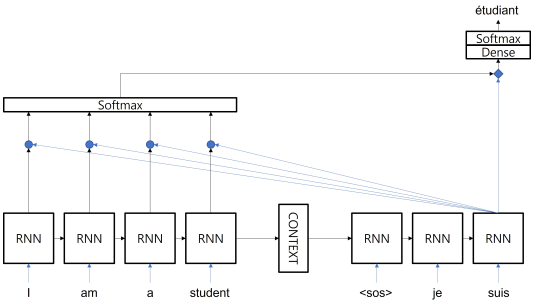


그림 3. Seq2Seq 모델에서의 Attention
Fig. 3. Attention in Seq2Seq model

태와 디코더의 현재 은닉 상태의 유사도를 확률 분포 형태로 표현하게 되고 이렇게 구한 확률값을 각각의 인코더의 은닉 상태에 곱하여 합하게 된다. 상기 어텐션 함수에서처럼 해당 값을 어텐션 값이라 하며 어텐션 값은 디코더의 은닉 상태와 결합(Concatenate)하여 하나의 벡터로 만든다. 해당 벡터는 최종 출력층의 가중치 행렬 및 Softmax 층을 거쳐 예측할 단어들의 확률을 구한다.

2.4 DETR(DEtection TRansformer)

DETR 모델[14]은 [그림 4]에서 볼 수 있듯이 3가지 요소로 구성되어 있다. 첫 번째로 이미지의 특징 추출을 위한 CNN (Convolutional Neural Network) backbone, 두 번째로 인코더 디코더 트랜스포머, 세 번째로 Detection을 위한 FFN(Feed Forward Networks)이다. Backbone으로는 ResNet을 사용하며 1x1 Convolution을 통해 트랜스포머의 은닉 차원(hidden dimension)으로 채널의 차원을 줄였다. 그리고 이미지를 1차원 평탄화하여 하나의 시퀀스를 생성하고 트랜스포머의 입력으로 사용하였다.

기존의 트랜스포머와 같이 포지셔널 인코딩(Positional Encoding)을 적용하는데 이미지에 대한 인코딩을 위해 2차원 형태의 포지셔널 인코딩을 진행한다. 2차원 형태의 포지셔널 인코딩은 x축 y축으로 나누어 인코딩하고 해당 인코딩 벡터를 합쳐서 최종 포지셔널 인코딩을 생성한다. 그리고 트랜스포머의 인코더와 디코더를 거쳐서 물체 최대 개수만큼 출력 임베딩이 생성되고 각각 독립적으로 FFN에 전달되어 물체의 종류와 위치에 대한 Bounding Box를 예측한다.

기존의 물체 인식 모델에서 사용하는 Set Prediction은 후처리 과정인 NMS (Non-Maximum Suppression), Anchor set 설계, Target Box를 Anchor에 할당하는 방법 등 간접적으로 문제를 해결했다. 상기 모델을 제시한 논문[14]은 직접 Set Prediction을 위한 트랜스포머 구조와 이분 매칭 손실 함수를 제안한다.

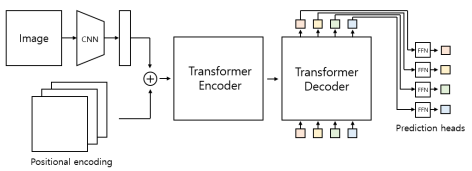


그림 4. DETR 구조[14]
Fig. 4. DETR Architecture

Ⅲ. 데이터 수집 환경 구성

3.1 RF 레이더 센서 및 모션 캡처 카메라 환경 구성

비가시 영역에서 탐지하고자 하는 물체에 대한 반사된 신호를 취득하기 위해 RF 레이더 센서 측정 시스템은 [그림 5]과 같다. 크게 레이더 신호를 생성하여 송·수신하는 RF 레이더 칩, 송신 신호를 증폭하기 위한 증폭기와 증폭기에 연결된 DC 파워 서플라이, MIMO 레이더 안테나 배열 구성을 위한 각각의 송·수신 스위치와 다수의 레이더 안테나로 구성되어 있다.

RF 레이더 칩은 Novelda 사의 임펄스 레이더 칩인 NVA-R631 모델을 이용하였다. 해당 레이더 칩은 0.45 ~ 3.55GHz 대역폭을 가지고 거리 분해능이 약 7.8mm이다. 증폭기는 Mini-Circuit 사의 ZHL-5W-422+ 모델로, 0.5 ~ 4.2GHz 주파수와 호환되며 RF 신호가 벽을 투과하기 위해 해당 대역폭의 주파수 신호들의 세기를 25dBm 만큼 증폭시킨다. 송·수신 스위치는 Mini-Circuit 사의 USB-1SP16T-83H 모델로, 호환 주파수가 1 ~ 8GHz이며 스위칭 딜레이는 5μs이다. 마지막으로 RF 레이더 안테나는 Novelda 사의 NVA-A03인 비발디 안테나를 사용하였고 1.3 ~ 4.4GHz의 주파수 대역을 지원하여 RF 레이더 칩의 신호와 호환되게 구성하였다.

레이더 안테나를 통해 MIMO 안테나 배열을 구성하여, 탐지하고자 하는 목표에 대한 반사 신호를 더 높은 해상도로 수집할 수 있다. 이는 송·수신 안테나 배치에 따라 해상도에 대한 차이가 존재한다. 본 논문에서는 [그림 6]과 같은 안테나 배치를 통해 MIMO 안테나 배열을 구성하였다. [그림 6]에서처럼 8개의 송신 안테나를 양 끝의 세로로 배치하였고, 8개의 수신 안테나를 가로로 배치하여 2차원 배열로 구성하였다. 그리고 레이더 안테나 간의 간격을 일정하게 유지하고 위치를 고정하기 위해 안테나 거치대를 사용하

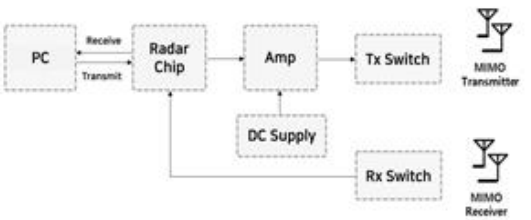


그림 5. RF 레이더 센서 구성
Fig. 5. RF Radar Sensor Architecture

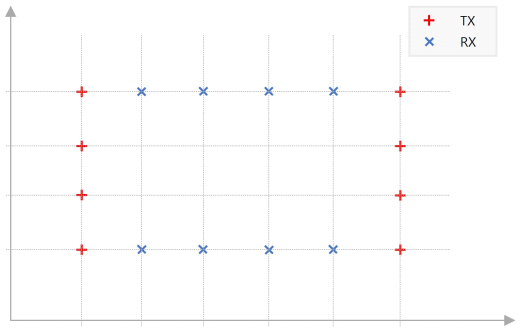


그림 6. 안테나 배치도
Fig. 6. Antenna Layout

여 구성하였다.

MIMO 안테나 배열을 통해 [그림 7]과 같이 가상 안테나 배열을 구할 수 있는데, 각각의 송·수신 안테나 간의 쌍을 지어 각 쌍을 통해 높은 세기의 RF 신호를 얻을 수 있는 가상의 위치를 특정할 수 있다. 이를 통해 탐지하고자 하는 목표의 해상도가 높게 물리적인 공간을 파악할 수 있으며, 데이터 수집에 대한 공간을 설정 및 조율할 수 있다.

객체의 공간 좌표를 나타내는 Label 데이터를 위해 모션 캡처 카메라를 [그림 8]과 같이 구성하여 물체에 대한 공간 좌표를 수집한다. 모션 캡처 카메라는 Opti-Track 사의 Flex 13 모션 캡처 카메라를 8대 설치하고, 동기화 디바이스를 통해 카메라 동기화를 한다. 동기화된 카메라를 이용하여 객체에 장착된 마커

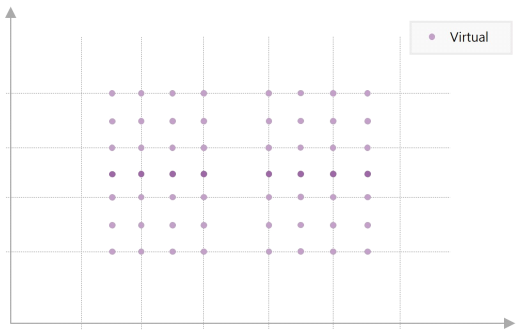


그림 7. MIMO 가상 안테나 배열
Fig. 7. MIMO Virtual Atenna Array

그림 8. 모션 캡처 카메라 구성
Fig. 8. Motion Capture Camera Settings

를 통해 객체의 공간 좌표 데이터를 수집한다.

3.2 학습 데이터셋 취득 공간 설정

학습 데이터셋을 수집하기 위해, 아래의 [그림 9]와 같이 모션 캡처 카메라를 배치하고, 수집한 Label 데이터의 원점 좌표와 레이더 안테나의 원점 좌표를 일치시켜 공간적 오차를 통해 발생하는 계산의 복잡성을 최소화한다.

객체가 움직일 수 있는 공간을 x축으로 $3m$, y축으로 $3m$, z축으로 $2m$로 설정하고, [그림 10]과 같이 $4.5cm$ 두께의 합판을 레이더 안테나와 객체 사이에 위치시켜 비가시 공간을 구현하고 데이터를 수집한다.

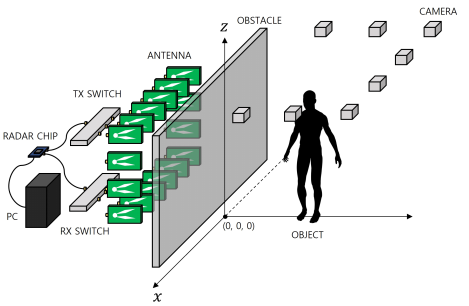


그림 9. 비가시 공간 구현도
Fig. 9. Realization of Invisible Space

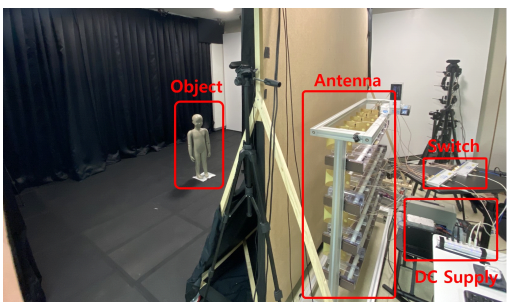


그림 10. 테스트베드 공간 배치
Fig. 10. Testbed space configuration

Ⅳ. 학습 모델 구성

학습 모델로는 오른쪽 상단의 [그림 11]의 트랜스포머 모델을 활용한다. 트랜스포머 모델은 레이더 신호의 1차원의 시퀀셜 데이터 특성과 각 송·수신 안테나 쌍의 거리 차이로 생기는 물체에 반사된 수신 신호의 특성 변화의 관계성을 파악하기 위해 사용한다. 또한, 트랜스포머는 기존의 Seq2Seq의 구조인 인코더-디코더의 구조를 가지지만, RNN을 사용하지 않고

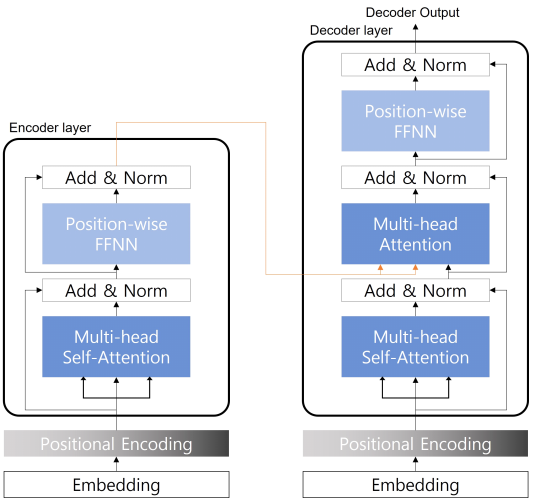


그림 11. 트랜스포머 구조
Fig. 11. Structure of Transformer

어텐션만으로 구현한 모델이다. RNN을 사용하지 않음으로써 RNN을 사용한 구조보다 상대적으로 가벼운 구조를 가지며, 이를 통해 컴퓨팅 자원 소모에 대한 부담이 적다.

4.1 특징 추출 및 레이더 신호 임베딩

샘플링된 레이더 신호의 특징을 추출하고 추출한 특징을 하나의 임베딩 벡터로 표현하기 위해 합성곱 인공신경망을 사용한다. 합성곱 인공신경망은 1차원 레이더 신호의 특징을 추출하기 위해 스트라이드를 (1, 2)로, 크기가 4인 1차원 커널인 (1, 4) 커널을 사용하고 출력 채널의 크기가 16인 합성곱 층을 4개를 쌓아 구성하였다. 구성한 합성곱 인공신경망을 통해 아래의 [그림 12]와 같이 송신 안테나와 수신 안테나 조합을 통해 얻은 1차원의 샘플링 신호의 차원을 줄이고 특징을 추출하고 채널 방향으로 임베딩을 수행한다. 이렇게 나온 인공신경망의 출력은 트랜스포머 입력으로 활용하는데, 각 안테나 쌍의 추출된 특징 간의 관계성 정보 학습을 위해 평탄화(Flatten)를 하여 입력 시퀀스 및 그에 따른 임베딩 벡터를 생성한다.

4.2 인코더 임베딩 벡터 및 포지셔널 인코딩

앞선 합성곱 인공신경망을 통해 특징 추출 및 임베딩을 수행한 입력 데이터의 형태는 임베딩 벡터의 크기를 d_{model}이라 할 때 트랜스포머 인코더의 입력으로 받을 임베딩 벡터의 크기는 $d_{model} = 16$이다. 그리고 레이더의 신호의 샘플 개수를 n_s, 각 송신 안테나와 수신 안테나의 개수를 n_{Tx}, n_{Rx}라 할 때 시퀀스의 길이는 $n_{Tx} \times n_{Rx} \times (\frac{n_s}{16} - 2)$이다.

트랜스포머는 앞선 내용처럼 RNN을 사용하지 않아 단어를 순차적으로 입력받아서 처리하지 못한다. 따라서 단어의 위치 정보를 다른 방식으로 처리하는데, 해당 정보를 각 단어의 임베딩 벡터에 위치 정보를 포함하는 포지셔널 인코딩을 더하여 모델의 입력으로 사용한다. 포지셔널 인코딩 벡터는 문장에서 단어의 위치별로 유일한 벡터값이어야 하고 서로 다른 길이의 문장에 대해 위치의 간격이 같아야 한다. 시퀀스에서 단어의 위치를 pos, 임베딩 벡터에서의 인덱스를 i라 할 때 포지셔널 인코딩의 수식은 다음과 같다.

$$PE_{(pos,2i)} = \sin(pos / 10000^{2i/d_{model}}) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos(pos / 10000^{2i/d_{model}}) \quad (2)$$

하나의 임베딩 벡터에 대해 사인과 코사인을 반복하여 표현하고 그 해당 값은 시퀀스에서 단어의 위치에 따라 일정하게 변하게 된다. 그리고 위의 수식에서 10000은 최대 시퀀스의 길이를 의미한다.

이렇게 구한 임베딩 벡터와 포지셔널 인코딩은 [그림 13]과 같이 더하여 트랜스포머의 입력으로 사용한다.

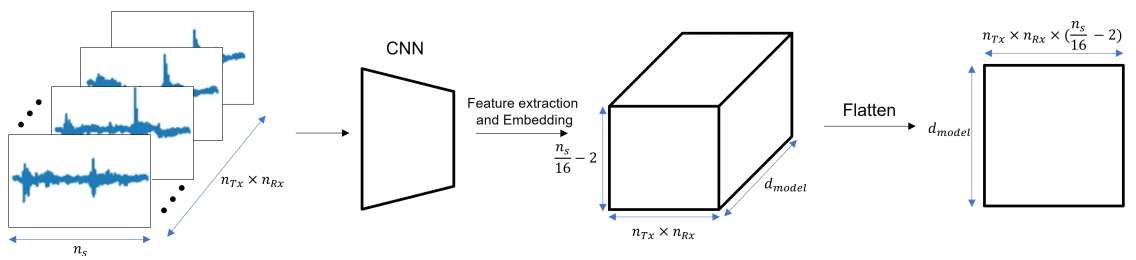


그림 12. 특징 추출 및 레이더 신호 임베딩
Fig. 12. Feature Extraction and Embedding of Radar Signal

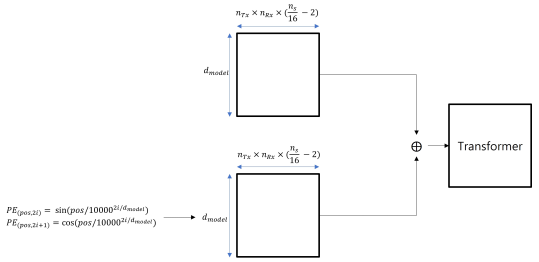


그림 13. 트랜스포머 인코더 입력 구성
Fig. 13. Input Structure of Transformer Encoder

4.3 인코더 멀티 헤드 셀프 어텐션

트랜스포머의 입력으로 받은 임베딩과 포지셔널 인코딩의 합은 멀티 헤드 셀프 어텐션 층을 거치는데, 여러 헤드로 나누어 헤드마다 병렬적으로 셀프 어텐션을 진행한다. 셀프 어텐션은 입력 시퀀스에 대해서 같은 시퀀스 내의 정보 간의 관계성을 학습하기 위해 사용하는데, Q(Query), K(Key), V(Value)를 동일한 값으로 사용하여 같은 시퀀스 내의 정보 간의 관계성을 학습한다.

헤드의 수를 n_{heads}라 할 때, 기존의 입력 행렬에 크기가 $(d_{\mathrm{model}}, (d_{\mathrm{model}}/n_{heads}))$인 가중치 행렬을 곱해 임베딩 차원을 헤드 수(n_{heads})로 나누어 [그림 14]와 같이 헤드마다 셀프 어텐션을 병렬적으로 학습할 수 있도록 처리한다. 이렇게 멀티 헤드 셀프 어텐션을 통해 위치 정보가 포함된 샘플링된 신호는 상기 가중치 행렬인 학습 가능 파라미터를 통해 $d_{\mathrm{model}}/n_{heads}$ 개수만큼 추출하여 추출한 신호 값에 대해 모든 안테나 쌍의 유사도를 구하게 된다. 유사도를 구하는 과정은 먼저 가중치 행렬을 곱하여 구한 Q와 K^T를 곱하고 Softmax 층을 거쳐 아래의 [그림 14]와 같이 어텐션 스코어를 구한다. 이렇게 구한 어텐션 스코어는 각 안테나 쌍의 유사도를 나타낸다. 안테나 쌍의 유사도는 추출한 안테나 쌍의 신호에 해당하는 V에 곱하여 어텐션 값을 구한다. 구한 어텐션 값은 결합(Concatenate)하여 인코더의 Position wise FFNN(Feed-Forward Neural Network)를 거쳐 인코더의 출력 및 디코더의 멀티 헤드 어텐션의 입력으로 사용된다.

4.4 디코더 임베딩 벡터 및 포지셔널 인코딩

디코더에서의 임베딩 벡터는 기존의 트랜스포머에서 Teacher Forcing을 위해 사용하는 임베딩 벡터와 다르게 정의한다. 실험 공간에서 물체 탐지를 위한 객체를 N개라 했을 때 객체 N개에 대한 임베딩의 학습 파라미터를 디코더의 입력으로 사용하는데 해당 파라미터는 객체의 위치 정보를 학습 가능하도록 한다. 이렇게 입력으로 받은 디코더에서 해당 객체 간의 위치 정보의 관계성을 인코더의 멀티 헤드 어텐션 과정과 동일하게 수행하여 구하고 해당 관계성 정보는 디코더의 멀티 헤드 어텐션 층을 거치게 된다. 해당 벡터를 Object Query[5]라 한다.

4.5 디코더 멀티 헤드 어텐션과 Prediction Head

디코더에서의 멀티 헤드 어텐션 층은 안테나 쌍의 특징 간의 관계성 및 위치 정보가 포함된 인코더의 출력을 K(Key)와 V(Value)로 활용하고 Object Query 행렬을 셀프 어텐션 층에 거친 결과를 Q(Query)로 사용한다. 그리고 정의된 Q, K, V를 통해 인코더의 멀티 헤드 어텐션 과정을 동일하게 거친다. 해당 과정은

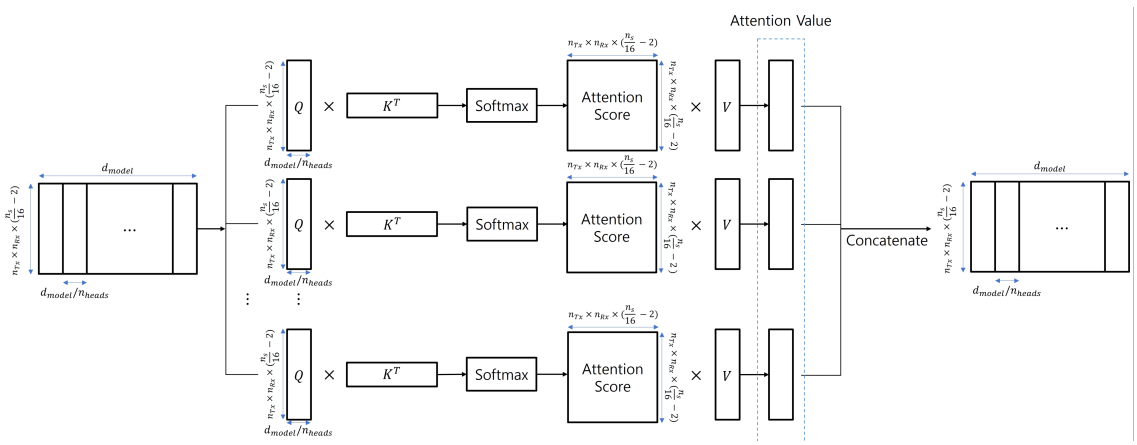


그림 14. 인코더 멀티헤드 셀프어텐션
Fig. 14. Multi-head Self-Attention

안테나 쌍의 신호 간의 관계성이 학습된 행렬과 각 객체의 관계성을 다시 한번 계산하고 계산된 관계성을 이용하여 어텐션 값을 계산한다.

계산된 어텐션 값은 인코더와 동일하게 FFNN 층을 거치고 디코더의 출력으로 나온다. 디코더의 출력은 최종적으로 Box를 예측하기 위한 Prediction Head를 거치는데 각 N개의 객체에 대해 독립적으로 거친다. Prediction Head는 3개의 FC-layer (Fully-Connected Layer)를 사용하였다. FC-layer의 마지막 층의 차원은 4이고 최종적으로 Sigmoid 함수를 통해 N개의 객체에 대해 정규화된 Box의 중심 좌표와 높이, 넓이를 예측한다.

V. 학습 결과

5.1 데이터셋 및 Label 데이터셋

학습을 위한 데이터는 RF 레이더 신호에 대해 38GHz의 샘플링 속도로 샘플링하였으며, 샘플링 개수는 1,024개이다.

그리고 스위치 제어를 통해 각 8개의 송신 안테나와 8개의 수신 안테나를 통해 얻을 수 있는 64개의 안테나 쌍의 신호에 대해 샘플링하여 수집하였다. 수집한 데이터의 형태는 [그림 15]와 같다.

이렇게 수집한 데이터는 트랜스포머에서 길이가 64이고 임베딩 벡터의 크기가 1024인 입력으로 사용하기 위해 차원 재설정을 통해 데이터에 대한 전처리 과정을 거친다.

Label 데이터는 상기 모션 캡처 카메라 환경 구성의 설명처럼 마커를 통해 물체의 삼차원 공간 좌표를 얻고 해당 좌표를 아래의 [그림 16]과 같이 Bird-Eye-View로 표현하여 xy 평면상의 좌표로 변환한다.

이렇게 변환한 좌표는 x축으로 최대, 최소의 좌표를 구하고 y축으로 최대, 최소의 좌표를 구해

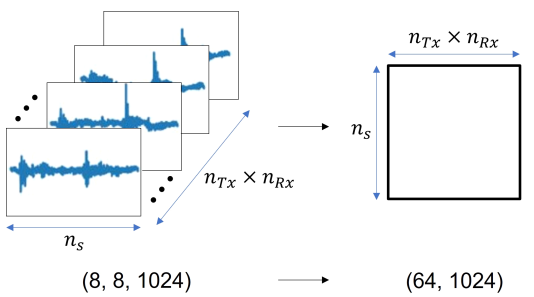


그림 15. 데이터 형태 및 사전 처리
Fig. 15. Shape of Data and Data Pre-processing

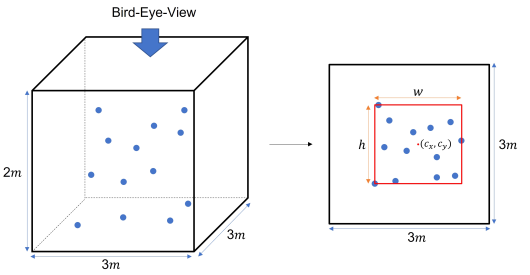


그림 16. Label 데이터 처리
Fig. 16. Processing Label Data

Bounding Box로 나타내고, 해당 Box의 중심 좌표 및 높이, 넓이를 구하여 해당 값을 Label 데이터로 활용한다.

탐지할 객체의 종류는 사람과 철제 의자로, 비가시 공간 내에 하나의 객체를 두고 이동시키며 측정하였다. 학습 데이터 및 Label 데이터는 사람에 대하여 6,500개, 의자에 대하여 4,500개를 수집하여 총 11,000개를 학습에 활용한다. 그리고 동일한 방법으로 테스트 데이터는 사람에 대하여 1,500개, 의자에 대하여 1,500개 수집하여 총 3,000개에 대하여 테스트하였다.

5.2 학습 손실 함수 정의

$$M_{cost} = c_{L1} L_{L1}(B_i, \widehat{B}_i) + c_{GIOU} L_{GIOU}(B_i, \widehat{B}_i) \quad (3)$$

먼저 제안한 모델은 한 번에 고정된 N개의 Bounding Box의 좌표를 예측한다. N은 비가시 공간 내의 최대 물체의 개수를 나타내는데, Target Box와 Prediction Box에 대해 손실 함수를 구하기 위해 이분 매칭(Bipartite Matching)을 생성한다. 매칭에 앞서 Target Box는 최대 물체의 개수에 맞게 N개가 되도록 패딩을 한다. 이렇게 패딩한 Target Box와 Prediction Box의 좌표를 이용하여 L1 norm과 GIoU를 구해 Cost로 사용한다. 이 Cost를 최소화하도록 헝가리안 알고리즘(Hungarian Algorithm)을 이용하여 둘을 이분 매칭한다. Cost를 수식으로 표현하면 다음과 같다.

이때 B_i와 $\widehat{B}_i$는 각각 Prediction Box와 Target Box의 좌표를 나타내고 c_{L1}과 c_{GIOU}는 상수이다. 이렇게 매칭된 인덱스를 이용하여 매칭된 Box를 구하고, 그 결과를 이용하여 전체 손실 함수를 구하게 된다. 이때 손실 함수에 대한 수식은 위의 매칭에 활용한 Cost 수식과 동일하게 사용한다. 해당 수식을 통해

나온 결과를 Batch 내에 존재하는 객체의 수로 정규화한다.

5.3 하이퍼파라미터 정의 및 테스트 결과

학습을 위한 Optimizer는 AdamW를 사용하였고 Learning Rate는 10^{-4}으로 설정하였다. 상기 학습 모델에서 트랜스포머의 인코더와 디코더 각각의 Layer 수는 모델의 복잡도를 파라미터 수를 줄여 낮추기 위해 1개로 정의하였고 트랜스포머의 FFNN 층의 은닉 차원은 2048로 설정하였다. 그리고 멀티 헤드 어텐션에서의 헤드 수는 8개로 정의하고 예측을 위한 최대 물체의 수는 100으로 설정하여 디코더에서 입력으로 받는 학습된 객체의 위치 정보의 다양성을 높여 학습을 진행하였다. 자세한 하이퍼파라미터 정의에 대한 표는 다음과 같다.

위의 학습 하이퍼파라미터에서 손실 함수를 위한 가중치인 c_{GIOU}와 c_{L1}은 그라디언트의 반영 비율에 맞춰 Box의 위치와 모양의 학습이 균형 있게 학습하도록 정의하였다.

이렇게 정의한 학습 하이퍼파라미터를 이용하여 11,000개의 학습 데이터에 대해 학습을 진행하고 학습한 결과에 대해 테스트 데이터로 테스트한 이미징 결과는 다음과 같다.

단일 물체에 대한 객체 인식 테스트 결과를 위해 Metrics를 정의하였는데, 예측 Box와 Target Box의 IoU를 구하고 사전 정의한 IoU Threshold를 이용하여 해당 값 이상이면 참으로 판단하여 정확도를 계산한다. 계산한 정확도의 결과는 다음과 같다.

레이더 신호의 특징 추출 및 임베딩을 위해 구성한 합성곱 신경망의 성능을 확인하기 위해 합성곱 신경망을 제거하여 샘플 신호의 차원을 임베딩 차원으로 정의하고 포지셔널 인코딩을 동일하게 더하여 트랜스포머의 입력으로 사용한다. 그리고 학습 하이퍼파라미

표 1. 학습 하이퍼파라미터
Table 1. Training Hyperparameter

Parameter	Value	Parameter	Value
Learning Rate	10^{-4}	Weight Decay	10^{-4}
Batch Size	4	Epoch	100
Encoder Layers	1	Decoder Layers	1
FFNN dimension	2048	dropout	0.1
Number of Heads (n_{heads})	8	Max Number of Objects (N)	100
GIOU Coefficient (c_{GIOU})	1	L1 Coefficient (c_{L1})	3

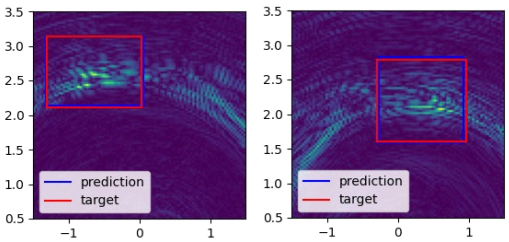


그림 17. 사람에 대한 레이더 신호 및 테스트 결과 이미징
Fig. 17. Radar and Test Result imaging of person

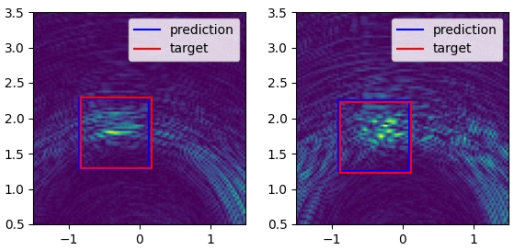


그림 18. 의자에 대한 레이더 신호 및 테스트 결과 이미징
Fig. 18. Radar and Test Result imaging of chair

터에서 예측 최대 객체 수(N)에 따른 성능 비교를 다른 하이퍼파라미터를 동일하게 정의하여 학습 후 그에 따른 정확도 결과의 차이를 통해 비교 분석한다. 비교 분석에 따른 결과 표는 다음과 같다.

위의 결과에 따라 합성곱 인공신경망과 최대 예측 객체 수 모두 객체 인식 성능에 영향을 끼친다. 합성곱 인공신경망을 추가함으로써 추가하지 않은 모델에 대해서 안테나 조합 간의 관계성 정보만을 이용하여 객체 인식을 수행한다면, 합성곱 인공신경망을 통해 특징 추출을 진행하고 안테나 조합의 추출된 특성 간

표 2. 테스트 정확도
Table 2. Accuracy in Test

IoU Threshold	0.7	0.5	0.3
Accuracy(%)	85.09	99.60	100

표 3. 합성곱 신경망과 최대 예측 객체 수에 따른 테스트 정확도 비교
Table 3. Comparison of Test Accuracy by Convolutional Neural Network and Max Number of Objects

IoU Threshold		0.7	0.5	0.3
Accuracy with CNN (%)	N = 100	85.09	99.60	100
	N = 2	19.62	63.70	90.92
Accuracy without CNN (%)	N = 100	4.20	20.55	46.83
	N = 2	31.25	65.74	78.32

의 관계성 정보 학습을 통해 객체 인식을 더 정확하게 수행할 수 있도록 한다. 그리고 합성곱 인공신경망을 이용한 객체 인식에서 최대 예측 객체 수가 증가할수록 객체의 위치 정보의 다양성이 증가하여 성능이 향상되는 것을 확인할 수 있다. 하지만 합성곱 인공신경망이 없는 경우 모델의 복잡도가 낮아 예측 경우의 수가 증가하여 성능이 떨어지는 것을 확인할 수 있다.

5.4 위치 추정 알고리즘 비교

트랜스포머를 이용하여 비가시 영역에서의 물체 위치 추정 알고리즘을 제안하였다. 위치 추정 알고리즘에서 삼차원 객체에 대해 2차원으로 변환하여 객체 인식 알고리즘을 구현하고 3차원 객체 인식 알고리즘에서 뛰어난 성능을 보이는 Complex YOLO[15]를 레이더 데이터를 학습하고 테스트를 통해 성능을 비교 및 분석한다.

Complex YOLO는 YOLO-v2[16] 기반으로 이미지에서의 물체 인식을 삼차원 객체 인식으로 개념을 확장한 알고리즘이다[15]. [15]에서 데이터로 라이다를 통해 얻은 포인트 클라우드를 2차원 RGB 이미지 매핑을 통해 얻은 이미지를 사용하였다. 먼저 학습을 위해 레이더 신호에 대한 전처리가 필요하다. 본 논문에서 사용한 데이터를 3차원 복셀(Voxel) 데이터로 변환하고 변환한 데이터를 2차원 RGB 이미지 매핑하여 학습에 사용하였다. 그리고 학습을 위한 Label 데이터는 기존의 데이터를 그대로 활용하였다. 이를 통해 본 논문에서 제안한 트랜스포머 기반 알고리즘의 성능과 Complex YOLO를 통한 성능에 대한 비교는 다음과 같다.

[표 4]와 같이 IoU Threshold에 따른 정확도에서 성능 차이가 큰 것을 확인할 수 있다. 레이더 신호를 그대로 사용하고 레이더 신호 간의 관계성 정보 학습을 통해 전처리 과정이 없이 객체 인식을 수행함으로써 레이더를 이용한 3차원 객체 인식에서의 가능성 및 성능을 확인할 수 있다.

표 4. 트랜스포머 기반 알고리즘 및 Complex YOLO 기반 알고리즘 테스트 정확도 비교
Table 4. Comparison of Test Accuracy by Algorithm with Transformer and Complex YOLO

IoU Threshold	0.7	0.5	0.3
Accuracy Transformer (%)	85.09	99.60	100
Accuracy Complex YOLO (%)	2.13	37.33	62.40

Ⅵ. 결 론

본 논문에서 UWB 레이더 칩과 다중 안테나 기술을 활용하여 레이더 신호를 통해 공간에 대한 정보를 수집하고, 수집한 정보를 합성곱 인공신경망을 통해 1차원 신호의 특징을 추출하고 추출한 특징을 기반으로 트랜스포머에 적용하여 단일 객체에 대한 Bird-Eye-View Box 예측을 이용하여 3차원 객체의 위치를 2차원 Box를 통한 추정 알고리즘을 제안하였다. 다중 안테나의 안테나 쌍의 거리 차이로 생기는 물체에 반사된 수신 레이더 신호의 특성 변화의 관계성 학습을 이용한 3차원 물체 인식 가능성 및 성능을 확인할 수 있었다.

현재 객체 인식은 3차원 위치를 바로 추정하지 않고 Bird-Eye-View를 통해 2차원 평면 상에 Bounding Box를 통한 위치 추정을 한다. 3차원 객체 인식을 위해 추가로 해당 모델을 발전시켜 해당 모델의 레이더 신호에 대한 더 정확한 3차원 공간 분석 가능성을 확인하고 이를 이용하여 포즈 추정을 위한 3차원 키포인트 인식 알고리즘을 개발 예정이다.

References

[1] C. R. Qi, et al., “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proc. IEEE CVPR*, pp. 652-660, Honolulu, Hawaii, Jul. 2017.

[2] B. Li, “3D fully convolutional network for vehicle detection in point cloud,” *2017 IEEE/RSJ Int. Conf. IROS*, pp. 1513-1518, Vancouver, BC, Canada, Sep. 2017.

[3] Y. Wang, et al., “Rodnet: Radar object detection using cross-modal supervision,” in *Proc. IEEE/CVF WACV*, pp. 504-513, Virtual, Jan. 2021.

[4] M. Meyer and G. Kuschk, “Deep learning based 3d object detection for automotive radar and camera,” *2019 16th EuRAD*, pp. 133-136, Paris, France, Oct. 2019.

[5] A. Vaswani, et al., “Attention is all you need,” *Advances in Neural Inf. Process. Syst.*, pp. 5998-6008, Long Beach, CA, USA, Dec. 2017.

[6] J. Li, Z. Zeng, J. Sun, and F. Liu, “Through-wall detection of human being’s movement by

UWB radar," in *IEEE Geosci. and Remote Sensing Lett.*, vol. 9, no. 6, pp. 1079-1083, Nov. 2012.

[7] H. Y. Son and K. W. Choi, "A study on pose estimation using radar signal and machine learning," *KICS Winter Conf. 2021*, pp. 1090-1091, Gangwon Province, Korea, Feb. 2021.

[8] K. Lu, et al., "Cascaded object detection networks for FMCW radars," *Signal, Image and Video Process.*, vol. 15, pp. 1731-1738, Apr. 2021.

[9] W. Kim, H. Cho, J. Kim, B. Kim, and S. Lee, "YOLO-based simultaneous target detection and classification in automotive FMCW radar systems," *Sensors 2020*, vol. 20, no. 10, May 2020.

[10] G. Zhang, H. Li, and F. Wenger, "Object detection and 3d estimation via an fmcw radar using a fully convolutional network," *IEEE ICASSP*, pp. 4487-4491, Virtual, May 2020.

[11] S. H. Chang, N. Mitsumoto, and J. W. Burdick, "An algorithm for UWB radar-based human detection," *2009 IEEE Radar Conf.*, pp. 1-6, Pasadena, CA, USA, May 2009.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, pp. 3104-3112, Montreal, Canada, 2014.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[14] N. Carion, et al., "End-to-end object detection with transformers," *Eur. Conf. Computer Vision and Pattern Recognition*, pp. 213-229, Virtual, Aug. 2020.

[15] M. Simony, et al., "Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds," in *Proc. ECCV Wkshps.*, Munich, Germany, Sep. 2018.

[16] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. IEEE Conf. CVPR*, pp. 7263-7271, Honolulu, Hawaii, Jul. 2017.

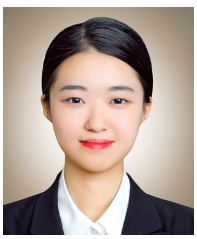
김 건 우 (Gon-Woo Kim)

2021년 2월 : 성균관대학교 전자전기공학과 학사 졸업
2021년 3월~현재 : 성균관대학교 전자전기컴퓨터공학과 석사과정
<관심분야> 레이더, 컴퓨터 비전, 머신러닝
[ORCID:0000-0003-1536-431X]

이 상 원 (Sang-Won Lee)

2021년 2월 : 성균관대학교 전자전기공학과 학사 졸업
2021년 3월~현재 : 성균관대학교 전자전기컴퓨터공학과 석사과정
<관심분야> 레이더, 머신러닝
[ORCID:0000-0001-8837-3255]

손 하 영 (Ha-Young Son)



2020년 2월 : 건국대학교 물리학과 학사 졸업
2020년 9월~현재 : 성균관대학교 전자전기컴퓨터공학과 석사과정
<관심분야> 컴퓨터 비전, 머신러닝
[ORCID:0000-0003-1870-8413]

최 계 원 (Kae-Won Choi)

2007년 8월 : 서울대학교 전기컴퓨터공학부 박사
2010년 9월~2016년 8월 : 서울과학기술대학교 컴퓨터공학과 조교수
2016년 9월~현재 : 성균관대학교 전기전자컴퓨터공학과 부교수
<관심분야> 무선통신, 무선전력전송
[ORCID:0000-0002-3680-1403]