한국ITS학회논문지 J. Korean Soc. Intell. Transp. Syst.

웨이블릿 변환을 활용한 효율적인 의미론적 분할 기술

Efficient Semantic Segmentation Using Wavelet-transform

안 택 현* · 최 정 단**

* 주저자 및 교신저자 : 한국전자통신연구원 초지능창의연구소 선임연구원
** 공저자 : 한국전자통신연구원 초지능창의연구소 책임연구원

Taeg-Hyun An* · Jeong Dan Choi*

* Electronics and Telecommunications Research Institute

† Corresponding author : Taeg-Hyun An, tekkeni@etri.re.kr

Vol. 23 No.5(2024)
October, 2024
pp.248~260

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2024.23.5.248

요 약

의미론적 영상 분할 기술은 오브젝트 검출과 더불어 자율주행 차량의 주변 환경 인식에 많이 사용되고 있다. 제한된 장비와 자원을 사용하는 자율주행 특성상 가볍고 빠른 네트워크가 선호되는데, 본 논문에서는 웨이블릿 변환을 활용하여 효율적인 의미론적 영상 분할을 하는 방법을 제안한다. 먼저 웨이블릿 변환을 사용하여 영상 데이터를 고주파, 저주파 성분으로 나누어 주고, 각각의 성분에 대하여 서로 다른 특징지도 추출을 하여 서로 다른 정보를 적합하게 합쳤다. 자율주행에 적합한 가벼운 네트워크를 베이스라인으로 Cityscapes 데이터 세트에 제안된 방식을 적용했을 때, 0.2%의 파라미터 증가를 통해 2.2% 성능향상을 달성했다. 이 같은 알고리즘을 활용하여 더욱더 안정적이고 정확한 주변 환경 인식에 적용되길 기대한다.

핵심어 : 의미론적 영상 분할, 웨이블릿 변환, 차량 주변 환경 인식, 자율주행

ABSTRACT

Semantic segmentation and object detection are widely used to perceive surrounding environment during autonomous driving. Owing to the nature of autonomous driving, which operates with limited resources and equipment, lightweight and fast networks are preferred. In this paper, we propose an efficient semantic segmentation algorithm using a wavelet transform. First, we apply the wavelet transform to separate high-frequency and low-frequency components from an input image. For each component, different feature maps are extracted, and the distinct information appropriately merged. When the proposed method was applied to the Cityscapes dataset using a lightweight network suitable for autonomous driving, a 2.2% performance improvement was achieved from a 0.2% parameter increase. We expect this algorithm can be applied to achieve more stable and accurate perceptions of the surrounding environment.

Key words : Semantic segmentation, Wavelet transform, Surrounding environment perception, Autonomous driving

Received 30 September 2024
Revised 11 October 2024
Accepted 17 October 2024

Ⅰ. 서 론

자율주행 차량에서의 주변 환경의 인식은 주행 경로의 계획 및 판단, 차량제어로 이어지는 자율주행의 중요한 과제 중 하나이다. 입력 영상이 들어오면 모든 화소에 대해 클래스 라벨을 부여하는 의미론적 분할은 주변 환경 인식에 사용될 수 있는 실용적인 기술이며, 이를 통해 주행 가능 영역, 차선, 주변 물체 등을 인식할 수 있다. 이러한 정보들이 정확하게 인식이 되면, 주행 경로 생성이나 물체 회피와 같은 이후의 과정들이 적합하게 동작하게 되어 안전하고 효율적인 자율주행이 가능하게 된다.

최근 컴퓨터비전이나 영상처리 관련 기술들은 심층신경망(Deep Neural Network, DNN)의 출현 이후 큰 발전을 이루어 왔다. 기존 다양한 형태의 수작업으로 설계된 특징(hand-crafted feature) 추출 방법들은, 2012년의 ImageNet competition에서 AlexNet(Krizhevsky et al., 2012) 이 보여준 성공적인 결과로부터 영향을 주기 시작한, 합성곱 신경망(Convolutional Neural Network, CNN) 구조를 위주로 한 심층신경망 방법들로 대체 되었으며, 이런 방법들은 영상 분류, 의미론적 분할, 객체 검출 및 탐지와 같은 분야에서 사용되는 다양한 벤치마크 셋에서 높은 성능을 나타내는 추세이다. 특히, 완전 합성곱 신경망(Fully Convolutional Network, FCN)(Long et al., 2015)은 기존 영상 분류에 사용되던 신경망을, 뒤 단에 존재하는 완전 연결 층(fully connected layers)을 제거하고 의미론적 영상 분할에 사용될 수 있게 신경망을 변형시킴으로써 이후의 의미론적 영상 분할의 방법들의 초석이 되었다.

이러한 발전에도, 실제 자율주행을 위해 사용될 의미론적 영상 분할 방법은 해결해야 할 과제들이 몇 가지 남아있다. 기본적인 성능을 나타내는 인식 정확도 이외에, 가장 중요하게 적용되어야 할 점은 실시간성 확보이다. 자율주행 차량은 매우 동적인 환경에서 동작하게 되는데, 예를 들어 10분의 1초 단위의 인식 지연이 발생하게 되면 주변 환경의 오차는 수 미터 단위로 날 수가 있고 이는 사고로 연결될 수 있다. 또한, 신경망의 연산 효율성 또한 매우 중요하다. 자율주행 차량은 한정된 자원과 장비를 사용하여 구동하여야 하기에 성능과 효율이 절충된 경량화된 모델을 개발이 요구된다. 또한, 주행환경은 각종 날씨와 다양한 조도 환경들이 존재하기에, 이러한 환경에서도 강건하게 동작하는 것이 요구된다.

본 논문에서는, 경량화 되어 있는 모델에서 복잡도와 연산시간이 많이 증가하지 않고도 쉽게 적용될 수 있는 방식으로 구성된 의미론적 분할 인공신경망 구조를 제안하고자 한다. 최근의 연구 중에는 DDRNet(Pan et al., 2022)과 같이 네트워크 구조설계를 함에 있어, 입력 특징지도를 특정 계층 이후부터 분기(branch) 시켜서 엣지 정보와 같은 디테일을 보존할 수 있는 detail branch와 그 이외의 함축적 정보를 많이 가지고 있는 semantic branch로 나누어서 특징지도를 생성한 뒤, 이를 합하여 최종 결과를 생성하는 모델 설계방식을 채택하는 방법들이 있다. 이러한 기존 방식들은 초기의 단순한 합성곱과 다운샘플링을 소수 수행하여 줄기가 되는 특징지도를 생성하고, 하나의 특징지도에서 특성이 다른 합성곱 계층들을 적용하여 semantic branch에는 함축적 정보를 잘 포함할 수 있게 하고, detail branch에는 디테일을 잘 보존할 수 있도록 유도하였다. 그리고, 중간중간 각각의 정보가 양방향으로 잘 섞일 수 있도록 합쳐주었다. 초기의 줄기가 되는 특징지도를 생성한 뒤에 잘 합쳐주는 앞선 방법과 달리 제안된 방법은, 애초에 입력 데이터를 함축적 정보가 더 잘 담길 수 있는 부분과 디테일이 더 잘 담길 수 있는 부분으로 분리하여 단순 합성곱과 다운샘플링을 소수 실행한 뒤 보다 빠른 단계에서 결합하고, 이후로는 전통적인 의미론적 분할 CNN 구조를 거쳐서 결과를 내도록 설계하였다. 이는, 이른 단계에서 입력 데이터를 잘 나누어서 효율적인 전처리를 통해 실시간 의미론적 분할 기술의 성능향상을 달성한 연구(An et al., 2023)의 방향성과도 부합한다. 입력 데이터의 정보의 효과적인 분리를 위해 웨이블릿 변환(wavelet transform)을 사용하였으며, 이를 통해 나온 고주파 성분과 저주파 성분을 분리하여 각각 detail branch와 semantic branch처럼 사용하였다. 특히 자율주행에 사용되는 가볍고 단순한 구조의 모델

네트워크 구조에서는, Cityscapes 데이터 세트에서 실험 결과 0.2%의 파라미터 숫자를 증가시켜서, mIoU 기준 2.2%의 유의미한 성능향상을 확보하였다.

본 논문의 구성은 다음과 같다. 먼저, 서론에 이어 II 장에서는 관련 연구에 관하여 기술하며, III 장에서는 제안된 의미론적 분할 방법과 추가적 센서인 LiDAR(Light Detection And Ranging, 빛을 통한 검출과 거리 측정) 센서와의 결합에 따른 확장 방법에 관해서 기술한다. 그다음으로, IV 장에서는 Cityscapes 데이터 세트와 KITTI-360 데이터 세트를 사용한 실험들을 통해 제안된 기술의 성능을 검증하였다. 마지막으로 V 장에서는 본 논문의 결론을 맺는다.

Ⅱ. 관련 연구

의미론적 영상분할은 컴퓨터비전 분야에서 폭넓게 연구됐으며, 최근에는 딥러닝을 활용하는 방식의 중요성이 부각되고 연구되는 추세이다. 이번 단원에서는 전통적 접근방식에서부터 최신 딥러닝 기반 아키텍처에 이르는, 이 분야의 핵심 연구들을 간략히 살펴본다.

1. 의미론적 분할(semantic segmentation)

딥러닝이 도입되기 전에는, 랜덤 포레스트나 서포트 벡터 머신(SVM)과 같은 전통적인 기계 학습 방법이 의미론적 분할에 사용되었다. 이러한 접근법은 색상, 질감, 에지 정보와 같은 수작업으로 설계된 특징을 사용하였으며 딥러닝 발전 이전 널리 사용되었다. 딥러닝, 특히 CNN의 도입은 다양한 컴퓨터비전 관련 작업이 그렇듯 의미론적 분할에서도 새로운 패러다임을 가져왔다. FCN(Long et al., 2015)은 의미론적 분할을 위한 최초의 딥러닝 모델 중 하나로, 기존의 분류(classification)에 사용되는 네트워크를 픽셀 단위의 클래스 라벨을 예측할 수 있도록 변형하였다. 이를 기반으로 한 SegNet(Badrinarayanan et al., 2017) 효율적인 인코더-디코더(encoder-decoder) 구조를 통해 픽셀 수준에서 정확한 예측을 가능하게 했고, 다중 스케일의 문맥 정보를 atrous convolution과 조건부 랜덤 필드를 결합하여 성능을 향상한 Deeplab(Chen et al., 2017)과 같은 방법들로부터 지속적인 발전이 있었다. 더욱 최근에 와서는 전통적인 CNN 기반 아키텍처 대신 이미지를 패치 단위로 나누고 언어학에서 사용되던 트랜스포머 모델을 사용하여 특징을 학습하는 방식의 Vision Transformer(ViT) (Dosovitskiy et al., 2021)가 출현하였으며, ViT를 의미론적 분할에 적용하게 되면서 더욱 정교한 전역 정보 학습이 가능하게 되었다(Zheng et al., 2021).

2. 자율주행에 사용될 의미론적 분할 모델

자율주행 차량에서 실시간으로 작동하는 의미론적 분할 모델은, 신속하고 정확한 의사결정을 위해 매우 중요하다. 이를 위해 ENet(Paszke et al., 2016)과 ERFNet(Romera et al., 2018)과 같은 경량화 된 네트워크 모델들이 개발되었다. 또한, MobileNet(Howard et al., 2017)과 ShuffleNet(Zhang et al., 2018)과 같은 경량화 된 아키텍처는 depthwise separable convolution을 사용하여 계산 복잡도를 크게 줄이면서도 우수한 성능을 유지하여, 모바일 및 임베디드 시스템에서 자주 사용되고 있다. 이러한 네트워크들은 자율주행 차량과 같은 제한된 자원 환경에서 높은 효율성을 보인다. 다양한 조명, 날씨, 가림 현상 등을 포함한 여러 환경에서 보여주는 강건함은 자율주행을 위한 의미론적 분할 모델의 또 다른 중요한 과제이다. 이를 해결하기 위해 연구자들은 데이터 증강, 도메인 적응(Hoffman et al., 2018) 등을 연구하고 있다.

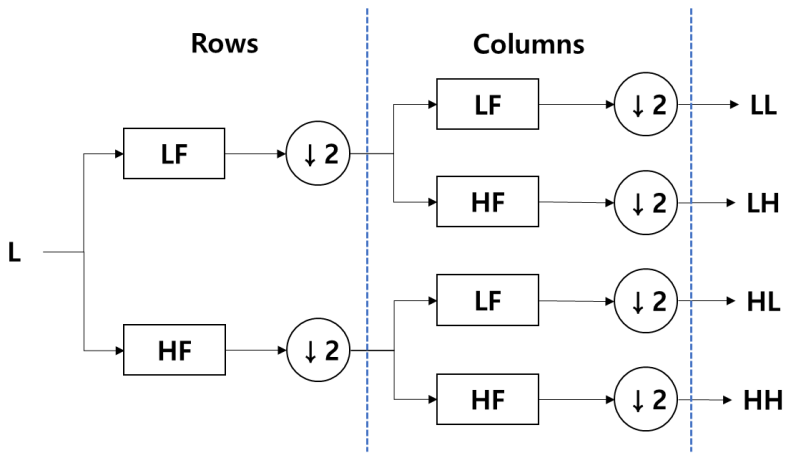
Ⅲ. 웨이블릿 변환을 이용한 심층신경망

이 장에서는 제안된 웨이블릿 변환을 이용한 의미론적 분할용 심층신경망에 관하여 기술한다. 먼저, 첫 번째 절에서는 웨이블릿 변환에 대하여 설명하고, 두 번째 절에서는 제안된 심층신경망 구조에 관하여 설명한다. 세 번째 절에서는 자율주행에서 다양하게 활용되는 LiDAR 센서 정보와의 결합을 통해, 제안된 방법의 확장 방법에 대해서도 알아본다.

1. 웨이블릿 변환(Wavelet transform)

웨이블릿 변환은 신호 처리에서, 시간 또는 공간 도메인에서의 다중 해상도 분석을 가능하게 하는 강력한 수학적 도구로, 주파수 영역에서 신호의 국부적 변화를 분석하는 데 사용된다. 이 변환은 신호를 저주파(전역적인 정보)와 고주파(세부적인 정보) 성분과 같은 부대역(subband) 성분으로 분리하여, 여러 스케일에서 다양한 수준의 세부 사항을 추출할 수 있어서, 에지 검출이나 영상 압축과 같은 곳에 활용되고 있다.

<Fig. 1>은 필터 뱅크를 통한 2D 웨이블릿 변환의 기본 구조를 나타낸다. 일반적으로 영상정보에 적용되는 2D 정보는 휘도 채널의 정보를 가지고 적용하기 때문에, 휘도를 뜻하는 L은 영상을 나타낸다. LF는 저역 통과 필터(Low-pass Filter), HF는 고역 통과 필터(High-pass Filter)를 뜻하며, 각 필터를 지나는 것은 합성곱 연산이 이루어진다.



<Fig. 1> Basic model of 2D wavelet transform

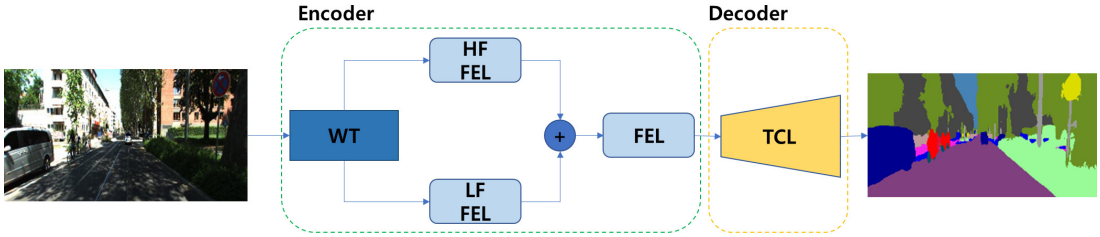
전체 과정을 살펴보자면, 원본 영상 정보 L이 들어오면, 행 방향으로 LF와 HF 필터링을 수행한 후, 열 방향으로 필터링을 수행한다. 각각의 필터링 후 결과 신호는 다운샘플링(보통 2배) 되어 주파수 성분을 분리해 준다. 따라서, LL 성분은 저주파 필터를 수평 및 수직으로 두 번 적용한 결과로, 영상의 전역적인 특징을 나타내고, LH 성분은 수평으로 저주파 필터를, 수직으로 고주파 필터를 적용한 결과로 수평 방향의 세부 정보를 포함하며, HL 성분은 수평으로 고주파 필터를, 수직으로 저주파 필터를 적용한 결과로 수직 방향의 세부 정보를 포함하고, 마지막으로 HH 성분은 고주파 필터를 두 번 적용하여, 대각선 방향의 세부 정보를 나타낸다. 본 논문에서는 웨이블릿 변환에서 가장 기본적인 필터 역할을 해 줄 수 있는 Haar wavelet 필터 뱅크를 사용하였다.

2. 네트워크 아키텍처

<Fig. 2>는 제안된 네트워크 구조를 나타낸다. WT는 웨이블릿 변환을 나타내며 R, G, B 각각의 채널에 대해 웨이블릿 변환을 수행하고, 각각의 성분에 아래 첨자로 채널을 명시하여 구분한다. 본 논문에서 저주파 성분은 <LL_R, LL_G, LL_B>로 행과 열 모두에 대해 저주파로 기저 성분을 의미하며, 고주파 성분은 <LH_R, LH_G, LH_B, HL_R, HL_G, HL_B, HH_R, HH_G, HH_B>으로 행과 열중에, 하나 이상에 대해 고주파로 세부 성분을 의미한다. 이후 각 고주파 성분 및 저주파 성분에 대해 특징추출 계층(Feature Extraction Layer, FEL)을 수행하게 된다. HF FEL은 고주파 성분의 특징추출 계층(High Frequency Feature Extraction Layer)을 나타내고, LF FEL은 저주파 성분의 특징추출 계층을 나타낸다. 그리고 나서는 각각에서 나온 특징지도를 합쳐준다. 제안된 방법은 애초에 입력 데이터를 함축적 정보가 더 잘 담길 수 있는 부분과 디테일이 더 잘 담길 수 있는 부분으로 분리하여 작업을 하였다. 대부분의 딥러닝을 사용하는 방법들이 그렇듯, 기존 Encoder-decoder 구조로 연결되는 단순한 형태의 의미론적 분할 네트워크 구조에서는 encoder 단에서 특징지도를 생성 함에 있어, 각각의 계층별로 어떠한 특징을 가지는 특징지도가 생성될지 알기 힘들다. DDRNet과 같은 방법들에서는 특정 분기 이후로, 두 가지 타입의 계층들을 만들어서 함축적 정보와 디테일 정보를 다룰 수 있게 네트워크 구조를 설계하고, 두 정보를 합침으로써 더 좋은 결과를 생성했는데, 제안된 방법에서는 DDRNet과 유사한 접근으로 두 가지 타입의 정보를 나누었는데, 웨이블릿 변환을 사용함으로써 초기 특징 자체를 함축적인 정보와 디테일 정보로 분리하여 정제한 뒤에 합쳐서, 더 좋은 결과를 생성하도록 하였다. 이후의 과정은 추가적인 특징추출 과정을 거쳐서 합쳐진 정보를 좀 더 정제한 다음, 결과 라벨지도를 생성하기 위한 decoder를 거친다.

웨이블릿 변환은 정해진 필터 뱅크를 사용하여 특징을 추출하고, 이들은 인공신경망의 특징추출과 달리 학습으로 결정되지 않는다. 따라서 정해진 필터에 의해 나오는 결과의 특징을 명확하게 알 수 있으며 여기서는 고주파 성분과 저주파 성분으로 나누는 역할로 사용되는데, 기존 인공신경망 구조의 분리 방법들에서는 두 가지 성분을 분리하기는 하지만 내부적으로 명확하게 성분의 특성을 지정해주지는 못하는 점이 있는 것을 보완해준다. 이 점은 기존 웨이블릿 변환을 사용한 방법들에서, 웨이블릿 변환의 단계별 고주파 성분을 사용해서 일반적 형태의 CNN에 추가적 특징으로 결합해주거나(Azimi et al., 2018), 디테일을 잘 복구하는 것에 초점이 맞추어져 있는 것과도(Zhao et al., 2021) 다르다고 볼 수 있다.

웨이블릿 변환은 LL 성분을 다시 하위 밴드로 분해하여, 단계적 구조를 가질 수 있으나 본 논문에서는 한 번의 단계만으로도 충분한 성능을 보일 수 있음을 확인하였다.



<Fig. 2> Proposed network architecture for semantic segmentation

<Table 1>에서는 각각의 계층별로 입출력 특징지도의 크기를 고려해 가며, 구체적인 네트워크 구조를 설명한다. 입력 데이터는 뒤의 실험에 활용될 Cityscapes(Cordts et al., 2016) 데이터 세트의 입력 영상을 절반으로 리사이즈 한 크기인 1024 x 512 크기가 기준이 된다. 먼저 Haar 필터뱅크를 이용해 RGB 각각의 채널에

대해 웨이블릿 변환을 수행하여, 입력을 고주파 성분과 저주파 성분으로 나눈다. 특징지도의 크기는 원래 크기에 비해 높이와 너비가 각각 절반이 되며, RGB 각각의 채널에 대해 저주파 성분과 고주파 성분의 채널 개수는 각각 1, 3이 되므로, 이들을 단순 연결(concatenate)하여 생성한 전체 저주파 성분과 고주파 성분 특징지도의 채널 수는 3, 9가 된다. 그리고 저주파 성분과 고주파 성분의 특징지도는 각각 LF, HF를 통과하며 다시 절반의 크기가 되는데, 이때 1x1 커널사이즈를 가진 합성곱(1x1 conv)을 사용하여 정보를 정제함과 동시에 결과 특징지도의 채널 수를 조절해주고, 스트라이드(stride)를 2로 하여 다운샘플링 효과를 가지는 3x3 커널 크기의 합성곱(downsampler block)을 통과시켜서 해당 기능을 수행한다. 구체적으로 1x1 conv와 downsampler block 에서는 각각 1x1 크기의 커널과, 3x3 크기의 스트라이드 2를 가진 커널을 이용하여 합성곱을 수행한 뒤, 배치 정규화와(batch normalization) 이어지는 REctified Linear Unit(ReLU) 활성화 함수가 사용된다. 이후에 사용되는 모든 합성곱에도 배치 정규화와 이어지는 ReLU 활성화 함수가 사용된다. 다음으로 특징지도 추출(FEL) 단계 직전에서는, 고주파 성분과 저주파 성분을 합쳐주는데 원소별 덧셈 연산을 사용한다. 특징지도 추출단계에서는 encoder의 기능을 수행하기 위해 많은 계층을 쌓아서 특징지도를 생성했다. 이를 위해, dilation이 없는 3x3 convolution과 이어지는 dilation이 있는 3x3 convolution을 연속해서 수행하는 3x3 conv를 여러 번 사용하게 된다. 처음에 dilation이 1인 3x3 conv를 5번 수행하였고, 이후에 downsampler block을 통과시켜서 특징지도의 크기를 축소하였다. 이어서 3x3 conv를 8번 수행하는데, 여기서는 보다 함축적이고 전역적인 정보를 포함하기 위해, 차례대로 2, 4, 8, 16, 2, 4, 8, 16 크기의 dilation을 가지고 차례대로 합성곱을 수행한다. Decoder에서는 총 N_C개의 클래스를 가지는, 원 영상의 크기와 일치하는 라벨 지도를 복원하게 되는데, encoder를 거치며 압축된 특징지도로부터 stride가 2인 Transposed convolution(T-convolution)과, 이어지는 dilation이 1인 3x3 conv들로 구성된 계층을(Transposed Convolution Layer, TCL) 통과시켜 최종 결과를 생성한다.

본 논문에서 제안하는 네트워크의 원형이 되는 모델은 ERFNet이다. 베이스라인 모델은 ERFNet에서 모델 합성곱 함수들과 세부 하이퍼 파라미터들을 바꾼 상태이며, 보다 구체적으로는 <Table 1>에서 WT, LF FEL, HF FEL을 대신하여 단지 두 개의 downsampler block로 구성된 계층을 가지게 되어 특징지도는 분리되지 않으며, 입력의 크기인 1024x512x3에서 512x256x16을 거쳐, 256x128x64의 크기인 특징지도를 가지게 된다.

<Table 1> Detailed architecture of the proposed semantic segmentation

		stage	input	operators	output
	WT		1024 x 512 x 3	Wavelet transform	512 x 256 x 3
					512 x 256 x 9
Encoder	LF FEL	L1	512 x 256 x 3	1x1 conv	512 x 256 x 16
		L2	512 x 256 x 16	Downsampler block	256 x 128 x 64
	HF FEL	H1	512 x 256 x 9	1x1 conv	512 x 256 x 16
		H2	512 x 256 x 16	Downsampler block	256 x 128 x 64
	FEL	F1-F5	256 x 128 x 64	3x3 conv (5 times)	256 x 128 x 64
		F6	256 x 128 x 64	Downsampler block	128 x 64 x 128
		F7-F14	128 x 64 x 128	3x3 conv (8 times)	128 x 64 x 128
Decoder	TCL	D1	128 x 64 x 128	T-convolution	256 x 128 x 64
		D2-D3	256 x 128 x 64	3x3 conv (2 times)	256 x 128 x 64
		D4	256 x 128 x 64	T-convolution	512 x 256 x 16
		D5-D6	512 x 256 x 16	3x3 conv (2 times)	512 x 256 x 16
		D7	512 x 256 x 16	T-convolution	1024 x 512 x N_C

이후의 실험들에서 제안된 방법과 베이스라인 모델은 같은 환경에서 학습된다. 베이스라인 모델 대비 제안된 모델은, downsampler block이 웨이블릿 변환으로 대체가 됨과 동시에 2개의 분기인 LF, HF로 나누어져서 특징지도를 생성하게 되고, 특징지도의 채널 수를 맞춰주기 위해 1x1 conv가 추가되었으며, 각 분기를 처리해줘야 하므로 해당 구간에서 2배의 연산이 필요하게 된 차이가 있다. 이후의 구간에서는, 특징지도를 합칠 때 단순한 원소별 덧셈 연산을 사용함으로써 특징지도의 채널 수 및 파라미터 개수가 베이스라인 모델과 같게 된다. 차이가 되는 부분의 영향은 크지 않아, 베이스라인 대비 전체 파라미터의 개수는 302.3만 개에서 303.0만 개로 0.3% 정도의 증가만 있다. 파라미터의 숫자는 python에서 pytorch 모델의 연산량과 파라미터 개수를 측정해주는, thop 패키지를 이용하여 측정되었다.

3. LiDAR 센서와의 확장

<Fig. 3>은 LiDAR 센서를 활용해 확장 시킨 제안된 네트워크이다. LiDAR 센서는 최근의 자율주행 기술에 있어 매우 중요한 센서이며, 물체를 인식하거나 거리를 측정하는 일 등에 사용된다. 대부분의 LiDAR 센서는 차량 주변 모든 방향에 점 구름(point cloud) 형태로 정보를 주며, 이를 통해 전방위적인 인식에 활용될 수 있도록 하는 연구들이 있다(Kang et al., 2021). 본 논문에서는 전방 카메라를 활용한 의미론적 분할을 다루고 있으며, 여기에서는 LiDAR 센서 정보를 가장 직관적으로 사용하는 방법의 하나인 점 구름으로 나타나는 정보를 영상에 투영하여 영상에 나타나는 부분의 정보만 활용하도록 한다. LiDAR 포인트를 카메라 영상에 투영하는 과정은 3D 공간에서 얻은 LiDAR 데이터를 2D 영상 좌표계로 변환하는 작업이며, LiDAR와 카메라의 각 고유 좌표계 간의 외부 행렬 및 내부 행렬(extrinsic and intrinsic matrix)들을 알고 있으면 변환할 수 있다. 외부 행렬(EM)은 LiDAR 좌표계에서 카메라 좌표계로의 변환을 담당하는 외부 행렬은 회전 행렬 R과 변환 벡터 T로 구성이 되어 있으며, 내부 행렬(IM)은 카메라 자체의 내부 파라미터로 카메라의 초점거리인 focal length(f_x, f_y)와 중심점(c_x, c_y) 등의 정보가 포함된다.

$$EM = [R \mid T]$$
$$IM = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

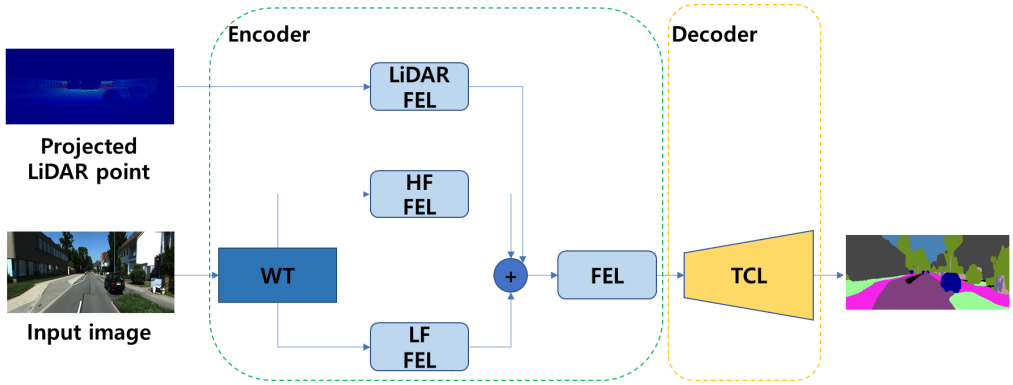
보다 구체적으로는 먼저 LiDAR에서 얻은 3D 점 구름 데이터를 카메라 좌표계로 변환해야 하며, 이때 외부행렬을 사용하여 LiDAR 좌표계의 위치인 [X_{LiDAR}, Y_{LiDAR}, Z_{LiDAR}]를 카메라 좌표계인 [X_{cam}, Y_{cam}, Z_{cam}]로 변환한다.

$$\begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \end{bmatrix} = R \begin{bmatrix} X_{LiDAR} \\ Y_{LiDAR} \\ Z_{LiDAR} \end{bmatrix} + T$$

카메라 좌표계로 변환된 3D 점 구름 데이터를 2D 영상 좌표계로 변환하기 위해서는 내부 행렬을 사용하여 변환한다.

$$\begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{Z_{cam}} \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \end{bmatrix} \begin{bmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \end{bmatrix}$$

여기서 u와 v는 영상 좌표에서의 화소 위치를 나타내며, 이는 최종적으로 LiDAR 점구름 데이터가 투영된 영상에서의 위치를 의미한다. LiDAR 데이터는 3차원의 좌표 정보, 강도(intensity)와 같은 정보를 가지게 되는데, 각각의 정보별로 영상 크기의 변환된 데이터로 가질 수 있게 된다. 예를 들어, 깊이와 강도만 가지고 온다면 2개의 채널을 가진 영상 크기의 데이터를 얻게 되는 것이다.



<Fig. 3> Network extension with a LiDAR sensor

<Table 2>에서는 LiDAR 특징추출 계층(LiDAR Feature Extraction Layer, LiDAR FEL)에 대한 구체적인 구조를 설명한다. 본 논문에서는, LiDAR 정보 중 깊이와 강도를 특징으로 사용하였으며, 그에 따라서 입력 채널 수는 2가 된다. 두 번의 downsampler block을 통과시켜서 특징지도의 크기를 축소하면서 채널 숫자를 늘려가며 HF FEL, LF FEL을 거쳐 나온 특징지도와 크기가 같도록 계층을 구성하였고, 원소별 덧셈 연산으로 특징지도들을 합쳤다. 합쳐진 특징지도의 크기는 LiDAR가 없는 경우와 비교하여 변하지 않았기에, 이후의 과정은 같다. LiDAR와 같이 영상에 투영될 수 있는 센서로부터의 정보들은 같은 방법으로 확장이 가능할 것이다. 영상에 투영된 LiDAR 점구름들의 깊이와 강도 정보는 기존 영상이 가진 RGB 정보에 비해, 모든 화소에 정보가 존재하지 않고 밀도가 희박하지만 3차원의 형상과 강도에 대한 추가적인 정보가 제공되는 것이기에, 영상만 사용하는 것에 비해서 성능향상이 될 것으로 기대된다.

<Table 2> Detailed architecture of the LiDAR feature extraction layer

		stage	input	operators	output
Encoder	LiDAR FEL	Li1	1024 x 512 x 2	Downsampler block	512 x 256 x 16
		Li2	512 x 256 x 16	Downsampler block	256 x 128 x 64

Ⅳ. 실 험

본 논문에서는 자율주행 차량에서 사용할만한 데이터 세트인 Cityscapes 벤치마크 데이터 세트를 사용하여 제안된 모델의 성능을 검증하였고, 추가로 LiDAR 센서와의 결합을 함께 보기 위해 KITTI360(Liao et al., 2022) 데이터 세트를 사용하였다.

1. 데이터 세트

Cityscapes 데이터 세트는 도시 환경에서의 의미론적 분할 및 객체 탐지를 위한 벤치마크 데이터 세트다. 독일의 다양한 도시에서 수집된 영상들로 구성되어 있으며, 각 영상에 대해 픽셀 단위의 주석(영상 분할 정답)이 제공된다. 총 5,000장의 2048 x 1024의 고해상도 영상으로 구성되어 있으며, 이 중 2,975장은 학습용, 500장은 검증용, 1,525장은 테스트용으로 구분되어 있다. 주석된 레이블은 자동차, 보행자, 건물 등과 같은 19개의 클래스로 구분된다. 본 논문에서는 정확한 의미론적 분할 성능을 평가하기 위해 Cityscapes 데이터 세트를 사용하였다.

KITTI-360 데이터 세트 자율주행 및 로봇 비전 연구를 위한 데이터 세트로, 주로 다중 모달(multi-modal) 데이터를 활용하여 도시 및 교외 환경에서의 객체 인식, 장면 이해 등을 평가하는 데 사용된다. 해당 데이터 세트는 다양한 2D, 3D 인식에 활용할 수 있도록, 스테레오 카메라, 어안렌즈 카메라와 LiDAR를 이용하여 얻은 데이터와 그에 해당하는 영상 분할, 객체 분할, 3D 상자 검출 등의 참값 데이터를 제공한다. 본 논문에서는 스테레오 카메라로 주어지는 영상 중 좌측 카메라에 해당하는 영상과, 그에 대응하는 참값 레이블, 같은 시간대에 저장된 LiDAR의 점 구름 데이터를 사용하였다. KITTI-360에서 제공되는 영상은 총 9개의 sequence로 이루어져 있으며, 영상의 해상도는 1408 x 384로 제공된다. 본 논문에서는 스테레오 영상이 페어로 존재하고, 그에 따른 LiDAR 정보도 함께 있는 61,186장의 영상을 데이터 세트로 사용하였다. 의미론적 분할 주석은 총 19개 클래스로 Cityscapes와 같으나, 데이터에 거의 존재하지 않는 bus와 train은 검증에서 제외하고 실험하였다.

2. 네트워크 학습

본 논문에서는 의미론적 분할 네트워크 중, 실시간 동작이 가능하여 자율주행에 활용될 수 있는 ERFNet을 기반으로 네트워크를 개량하여 학습하였다. 학습환경으로는 Ubuntu 22.04 LTS에서 pytorch를 사용하였으며, 개발용 PC 사양은 intel i-9 10980XE, NVIDIA-RTX 3090, 128GB RAM이다.

Cityscapes와 KITTI-360 각각의 학습에 있어서 공통적으로 최적화 함수는 ADAM을 사용하였고, 이와 관련된 beta-1은 0.9, beta-2는 0.999로 설정하였으며, learning rate는 Cityscapes의 경우 0.001, KITTI-360은 0.01로 설정하였다(Kingma and Ba, 2015). 학습할 때 사용된 loss 함수는 Cross-entropy이며, 500 epoch 동안 학습을 수행하였다.

Cityscapes의 경우, 학습용 데이터와 검증용 데이터를 이용하여 학습 및 실험 결과를 관찰하였다. KITTI-360의 경우, 총 9개의 영상 시퀀스 중에 6개를(총 45,108장, 시퀀스 넘버 0, 2, 3, 4, 5, 6) 학습에 사용하고, 3개의 시퀀스를(총 16,060장, 시퀀스 넘버 7, 9, 10) 검증에 사용하였다. 시퀀스 영상의 특성상 중복되는 장면이 매우 많기에, 최종적으로는 학습과 검증에 사용되는 영상을 10개 단위로 샘플링하여, 학습과 검증에 사용되는 시간을 줄여서 실험하였다.

3. 실험 결과 및 평가

본 연구에서 제안한 모델의 성능을 평가하기 위해 mIoU(mean Intersection over Union) 와 네트워크를 구성하는 파라미터의 개수를 주요 지표로 사용하였다. mIoU는 의미론적 분할 작업에서 널리 사용되는 평가 지표 중 하나로, 각 클래스에 대한 예측의 정확도를 픽셀 단위로 평가하고, 모든 클래스에 대해 그 평균을

계산한다. IoU는 mIoU를 계산하기 위해 클래스별로 사용되는 정확도로, 각각의 클래스에 대해 모델이 예측한 레이블 영역과 정답 레이블 간의 겹치는 부분과 전체 영역 간의 비율을 측정한다. 예측과 정답이 겹치는 부분은 TP(True Positive), 전체 영역이라고 함은 TP와 FP(False Positive), FN(False Negative) 모두를 합한 영역이 된다:

$$IoU = \frac{TP}{TP + FP + FN}$$

$$mIoU = \frac{1}{N_c}\sum_{c=1}^{N_c} IoU_c$$

또한, 네트워크가 효율적으로 구성되었는지 알아보기 위해 파라미터 개수와 연산량이 얼마나 증가하는지를 동시에 관찰하였으며, 이를 위해서 python에서 pytorch 모델의 연산량(MACs, Multiply-ACcumulate operations)과 파라미터 개수를 측정해주는, thop 패키지를 이용하였다.

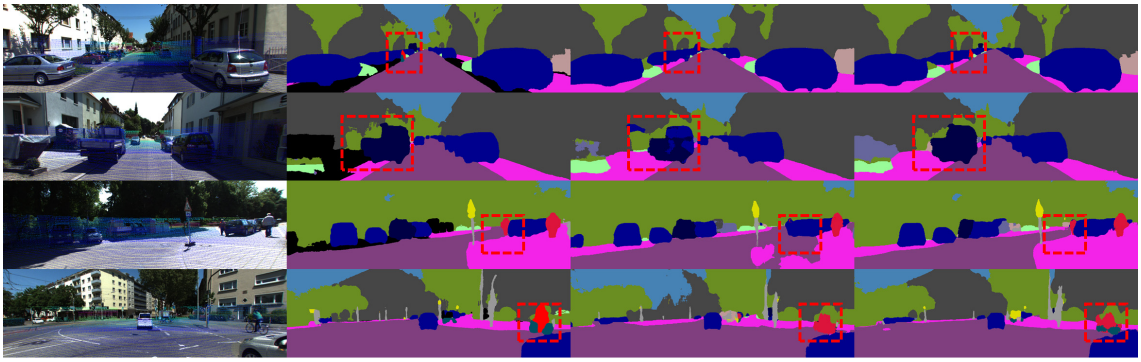
<Table 3> quantitative results with Cityscapes dataset

	# of parameters	MACs	mIoU
baseline	3,023K	42.54G	0.7085
proposed	3,030K	42.76G	0.7305

<Table 3>과, <Table 4>는 각각 Cityscapes와 KITTI-360에서 실험한 결과를 정량적으로 나타내고 있다. 두 실험에서 모두 1%가 되지 않는 파라미터 숫자와 연산량의 증가임에도 불구하고, 성능개선이 명확하게 이루어지는 것을 확인할 수 있다. Cityscapes 데이터 세트로 실험한 경우, 파라미터의 숫자는 3,023K에서 3,030K로 0.2 % 만큼 증가하였고, 연산량의 경우 42.54G에서 42.76G로 0.5%만큼 증가하였다. 이를 통해 성능향상은 70.85%에서 73.05%로, 2.2% 증가하는 것을 확인할 수 있다. KITTI-360의 경우에는 LiDAR로 확장한 모델(proposed-LiE)이 존재하기 때문에, 두 가지 형태의 모델로 나타나는데, 두 모델 모두 0.5% 이하의 파라미터 수 증가와 1.3% 이하의 연산량 증가를 통해서, 각각 2.21%(proposed), 3.49%(prposed-LiE)의 성능이 향상되는 것을 확인할 수 있다. 파라미터 및 연산량의 차이는 baseline 모델의 초기 계층에 존재하는 1개의 downsampler block이 2개의 wavelet transform과 이어지는 1x1 conv로 대체 되면서 발생하는데 그로 인한 차이가 1% 내외가 되겠고, 발생하는 성능향상은 2% 이상이므로 경량화가 중요한 자율주행용 모델에서는 합리적인 비용(cost)이라 할 수 있다.

<Table 4> quantitative results with KITTI-360 dataset

	# of parameters	MACs	mIoU
baseline	3,023K	42.86G	0.5442
proposed	3,030K	43,11G	0.5663
proposed-LiE	3,037K	43.39G	0.5791



<Fig. 4> Qualitative result with KITTI-360 dataset. (a) original image with LiDAR point cloud overlaid (b) Ground truth. The Black area is void. (c) result with baseline. (d) result with proposed-LiE.

<Fig. 4>는 실험한 결과를 정성적으로 나타내고 있다. 왼쪽부터 차례대로 LiDAR 점 구름을 투영해놓은 원영상, 참값 라벨, 베이스라인 네트워크로부터의 결과, 제안된 네트워크로부터의 결과(proposee-LiE)이다. 붉은색 점선 상자는 주로 차이가 나는 부분을 표시해놓은 것인데, 주로 물체 영역이다. 이를 통해 전체적으로 결과가 좋아지지만, LiDAR로부터 얻어내는 특징값인 깊이 정보와 강도를 통하여 물체 영역이 더 잘 나타나는 것을 관찰할 수 있다.

V. 결 론

본 논문에서는 경량화 되어 있는 모델에서 복잡도와 연산시간이 많이 증가하지 않고도 쉽게 적용될 수 있는 방식으로 구성된 의미론적 분할 인공신경망 구조를 제안하였다. 제안된 방법은 웨이블릿 변환을 이용하여 획득한 고주파, 저주파 성분을 통하여 초기 특징지도 자체를 함축적인 정보와 디테일 정보로 분리하여 정제한 뒤에 합침으로써 더 좋은 결과를 생성하도록 하였으며, 이후로는 전통적인 의미론적 분할 네트워크 구조를 거쳐서 결과를 내도록 설계함으로써, 기존의 의미론적 분할 네트워크 대비 복잡도가 많이 증가하지 않을 수 있게 하였으며, LiDAR와 같이 영상에 투영할 수 있는 센서들을 이용한 확장 방법에 대해서 보였다. 제안된 접근방식을 자율주행에 활용될 수 있는 Cityscapes 데이터 세트와 KITTI-360 데이터 세트에 각각 적용해본 결과, 자원이나 시간이 제한된 환경에서 효율적인 성능향상을 이룰 수 있음을 확인하였다. 향후, 다양한 형태의 네트워크를 개량하고 다양한 형태의 특징지도 결합 방식을 적용해봄으로써, 자율주행에 널리 활용할 수 있을 것으로 기대한다.

ACKNOWLEDGEMENTS

본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 RS-2021-KA161756, 과제명: 실시간 수요대응 자율주행 대중교통 모빌리티 서비스 기술 개발)

REFERENCES

An, T. H., Kang, J. and Min, K. W.(2023), “Network adaptation for color image semantic segmentation”, *IET Image Processing*, vol. 17, no. 10, pp.2972-2983.

Antonini, M., Barlaud, M., Mathieu, P. and Daubechies, I.(1992), “Image coding using wavelet transform”, *IEEE Trans. Image Processing*, vol. 1, no. 2, pp.205-220.

Azimi, S. M., Fischer, P., Körner, M. and Reinartz, P.(2018), “Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp.2920-2938.

Badrinarayanan, V., Kendall, A. and Cipolla, R.(2017), “SegNet: A deep convolutional encoder-decoder architecture for image segmentation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp.2481-2495.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L.(2017), “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp.834-848.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R. and Schiele, B.(2016), “The cityscapes dataset for semantic urban scene understanding”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3213-3223.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. and Houlsby, N.(2021), “An image is worth 16x16 words: Transformers for image recognition at scale”, *International Conference on Learning Representations (ICLR)*.

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K. and Darrell, T.(2018), “Cycada: Cycle-consistent adversarial domain adaptation”, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp.1989-1998.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T. and Adam, H.(2017), “Mobilenets: Efficient convolutional neural networks for mobile vision applications”, *arXiv preprint arXiv:1704.04861*.

Kang, J., Han, S. J., Kim, N. and Min, K. W.(2021), “ETLi: Efficiently annotated traffic LiDAR dataset using incremental and suggestive annotation”, *ETRI Journal*, vol. 43, no. 4, pp.630-639.

Kingma D. P. and Ba J. L.(2015), “ADAM: A method for stochastic optimization”, *in Proc. third International Conference on Learning Representations (ICLR)*, San Diego, California, pp.1-15.

Krizhevsky, A., Sutskever, I. and Hinton, G. E.(2012), “ImageNet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, vol. 25.

Liao, Y., Xie, J. and Geiger, A.(2022), “Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp.3292-3310.

Long, J., Shelhamer, E. and Darrell, T.(2015), “Fully convolutional networks for semantic segmentation”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3431-3440.

Pan, H., Hong, Y., Sun, W. and Jia, Y.(2022), “Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp.3448-3460.

Paszke, A., Chaurasia, A., Kim, S. and Culurciello, E.(2016), “ENet: A deep neural network architecture for real-time semantic segmentation”, *arXiv preprint arXiv:1606.02147*.

Romera, E., Alvarez, J. M., Bergasa, L. M. and Arroyo, R.(2018), “ERFNet: Efficient residual factorized convnet for real-time semantic segmentation”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp.263-272.

Zhang, X., Zhou, X., Lin, M. and Sun, J.(2018), “Shufflenet: An extremely efficient convolutional neural network for mobile devices”, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6848-6856.

Zhao, C., Xia, B., Chen, W., Guo, L., Du, J., Wang, T. and Lei, B.(2021), “Multi-scale wavelet network algorithm for pediatric echocardiographic segmentation via hierarchical feature guided fusion”, *Applied Soft Computing*, vol. 107, 107386.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y. and Yu, G.(2021), “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6881-6890.