Transactions of KSAE, Vol. 33, No. 11, pp.957-965 (November, 2025)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2025.33.11.957

서브-모달리티 주의 기반 효율적인 의미론적 분할 기술

안 택 현* · 민 경 욱 · 최 정 단

한국전자통신연구원, AI로봇 연구본부

Efficient Semantic Segmentation Based on Sub-Modality Attention

Taeg-Hyun An* · Kyoungwook Min · Jeong Dan Choi

Artificial Intelligence Robot Research Division, Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Korea

(Received 7 July 2025 / Revised 1 August 2025 / Accepted 4 August 2025)

Abstract : Semantic segmentation plays a crucial role in autonomous driving by assigning pixel-wise labels to images. Traditional convolutional neural networks (CNNs) based semantic segmentation approaches incorporate conventional down-sampling layers in the initial stage to enhance computational efficiency. The feature maps in the initial layers of a CNN are more effective when individual channels capture diverse and complementary information. Thus, this work introduced a sub-modality attention network that explicitly separates high-frequency and low-frequency components, focused on integrating separated pieces of information, allowing them to complement each other's deficiencies at the early feature extraction stage. Our results demonstrate that deepening CNNs is not the only path to performance improvement—incorporating handcrafted priors, such as the Wavelet transform, can also yield significant gains.

Key words : Semantic segmentation(의미론적 영상 분할), Real-time(실시간), Sub-modality(서브-모달리티), Attention(주의), Wavelet transform(웨이블릿 변환)

1. 서 론

자율주행 차량에서의 주변 환경 인식은 이후의 판단, 제어로 이어지는 전체 자율주행 시스템에서 핵심적인 역할을 수행한다. 특히 입력 영상의 각 화소에 의미 있는 범주(Label)를 할당하는 의미론적 분할(Semantic segmentation)은 주행 가능 영역 탐지, 차선 인식, 보행자나 차량 등의 객체 검출 등 다양한 인지 작업에 활용되는 실용적인 기술이다. 이러한 정밀한 인식은 주행 경로 생성, 장애물 회피 등의 후속 모듈에 직접적인 영향을 주어, 안전하고 효율적인 자율주행을 가능하게 한다.

전통적인 의미론적 분할 기법들은 서포트 벡터 머신(SVM), 랜덤 포레스트(Random forest) 같은 기계학습 기법과 수작업으로 정의된 특징(Handcrafted feature)에 기반해 왔다. 그러나 심층 신경망(Deep Neural Network, DNN)의 등장 이후, 모델이 입력 영상으로부터 계층적인 특징 표현을 스스로 학습할 수 있게 되면서 의미론적 분할 분야는 급격한 발전을 이룩하였다. 특히 완전 합성곱 신경망(Fully Convolutional Network, FCN)은 기존 영상 분류 모델을 분할 문제에 맞게 변형하여, 이후 다양한 신경망 기반 분할 모델의 기반이 되었다.

자율주행과 같은 실시간 응용을 위한 의미론적 분할에는 여전히 몇 가지 기술적 과제가 존재한다. 높은 인식 정확도는 물론, 다른 인식이나 판단, 그리고 제어용 작업들과 함께 작동하기 위하여 연산 효율성을 함께 만족해야 한다. 차량 내부의 한정된 연산 자원과 에너지 제약을 고려하면, 효율적인 네트워크 설계가 필수적이다.

본 연구에서는 입력 영상을 웨이블릿 변환(Wavelet transform)을 사용해 저주파(LF: Low Frequency)와 고주파(HF: High-Frequency)라는 서브-모달리티(Sub-modality)로 분해하고, 요소별로 특징지도를 추출해 준 다음, 각 요소의 특징지도를 서브-모달리티 주의(Sub-modality attention)로 적절히 결합하는 의미론적 분할 네트워크 구

*Corresponding author, E-mail: tekkeni@etri.re.kr

조를 제안한다. 두 성분으로 나누면서 초기 고해상도 연산 부분을 웨이블릿 변환을 포함한 가벼운 연산으로 변환하였고, 이를 통해 가능해진 병렬처리를 통한 속도향상뿐만 아니라, 의미 있는 정확도 향상을 달성할 수 있다. 기존 방식들은 주로 학습에 의존하는 계층들을 어떻게 만들지 위주였기에 각 계층이 어떤 역할을 하는지 알아보려면 별도의 시각화와 같은 작업을 수행해주어야 했지만, 제안된 방법은 웨이블릿 변환을 포함한 초기 계층에서는 한정적으로 역할을 제공해 주었으며 그로 인한 성능 향상을 확인할 수 있었다. 이를 통해, 네트워크 구조를 잘 만드는 것뿐만 아니라 입력을 어떻게 사용하는지도 중요한 것을 보였다고 할 수 있겠다.

제안된 방법은 자율주행 분야에서 활용되는 Cityscapes 데이터 세트를 기반으로, 영상 1장당 처리시간이 8.3 ms였던 기존의 경량화 모델을 개선하여 연산시간을 1.1 ms 추가로 단축했을 뿐만 아니라, mIoU 기준 성능을 70.85 %에서 74.72 %로 끌어올려 3.87 %의 유의미한 향상을 달성하였고, 단순 서브-모달리티 특징 결합방식과 비교해서도 성능 향상이 있음을 확인했다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련된 기존 연구 들을 정리하고, Ⅲ장에서는 제안하는 네트워크 구조와 서브-모달리티 주의 기반 접근법을 제시한다. Ⅳ장에서는 실험 결과를 통해 제안 방법의 성능을 검증하고, 마지막으로 Ⅴ장에서 본 연구의 결론과 향후 연구 방향을 논의한다.

2. 관련 연구

의미론적 분할(Semantic segmentation)은 컴퓨터 비전 분야에서 중요한 과제로, 이미지 내의 각 화소(Pixel)에 범주(Label)를 할당함으로써 객체 인식보다 더욱 정밀한 장면 이해를 가능하게 한다. 본 장에서는 기존 의미론적 분할 기술의 발전과 주의 메커니즘의 적용에 관한 연구 동향을 다룬다.

2.1 의미론적 분할 기술의 발전

의미론적 분할(Semantic segmentation)은 영상 내 각 픽셀에 대해 고유한 의미가 있는 레이블을 할당하는 과제로, 딥러닝 기술의 도입과 함께 괄목할 만한 발전을 이루었다. 특히 CNN 기반의 아키텍처를 중심으로 특징 추출과 공간정보 이해의 정밀도가 크게 향상되었다.

인코더-디코더 기반 모델은 이러한 발전의 핵심 구조 중 하나이다. U-Net[1]은 인코더를 통해 다양한 스케일의 특징을 추출하고, 디코더에서 이를 복원하며 정밀한 경계 정보를 복구하는 데 효과적인 구조를 제시하였다. FCN[2]은 분류용 신경망을 픽셀 수준의 예측에 맞게 변형하고, 스킵 연결을 도입하여 해상도 손실을 최소화하면서도 정확한 분할을 수행하였다. SegNet[3]은 풀링 인덱스를 활용하여 디코더에서 효율적인 업샘플링을 수행함으로써 연산량을 줄이면서도 세부 공간정보를 보존하였다.

다중 스케일 문맥 정보를 활용하는 기법들도 성능 향상에 중요한 역할을 해왔다. Atrous convolution(또는 Dilated convolution)은 합성곱 커널 간의 간격을 조절하여 넓은 수용 영역을 확보하면서도 계산량을 유지하는 방식이다. DeepLab[4] 시리즈의 ASPP(Atrous Spatial Pyramid Pooling) 모듈은 서로 다른 팽창률을 가진 필터를 병렬로 적용함으로써 다양한 스케일의 문맥 정보를 효과적으로 수용할 수 있도록 한다. PSPNet[5]은 피라미드 구조를 활용한 평균 풀링을 통해 전역 문맥 정보를 학습하고, 이를 결합함으로써 보다 안정적인 분할 결과를 도출한다.

자율주행은 계산 자원이 제한된 응용 분야이며, 경량화 및 실시간 처리 모델은 필수적이다. ERFNet[6]은 Dilated convolution, 1D factorized convolution, 그리고 잔차 연결을 조합하여 연산량과 성능 간 균형을 달성하였다. BiSeNet[7]과 DDRNet[8]은 이중 분기 구조를 통해 국소 정보와 전역 문맥을 동시에 포착하여 빠르면서도 정확한 분할을 가능하게 하였고, STDC[9]는 단순화된 네트워크 구조를 통해 실시간 응답성을 구현하였다. 이외에도, 다양한 환경에 적용하기 위해 도메인 변환[10]을 적용한다거나, 주차 구획 검출[11]에 사용하기도 하고, 깊이 정보를 얻기 위해 자율주행에서 사용되는 LiDAR 센서와의 결합을 통해 공간정보를 얻거나, 더 안정적인 결과를 얻는 것에 사용되기도 하며, LiDAR센서의 정보 자체로 의미론적 분할을 수행하기도 한다.[12-14]

2.2 Attention 기반 의미론적 분할 모델

딥러닝 모델에서의 Attention 메커니즘은 특징 선택(Feature selection) 능력을 향상함으로써, 모델이 보다 중요한 영역에 집중할 수 있도록 한다. 이는 의미론적 분할에서도 핵심적인 역할을 하며, 다양한 Attention 기반 구조들이 제안되고 있다.

SENet[15]은 각 채널의 중요도를 학습하여 채널 간 가중치를 조절함으로써 특징 표현을 향상시킨다. ECA[16]는 복잡한 연산 없이 채널 간 상관성을 동적으로 반영함으로써 SENet 보다 가벼운 구조를 유지하며 유사한 성능을 제공한다.

CBAM[17]은 채널 주의와 공간주의를 결합하여, 입력 특징지도의 전역 문맥 정보를 활용한 보정이 가능하다. 이는 특히 복잡한 배경이나 물체 간 경계가 모호한 환경에서 분할 성능을 개선하는 데 효과적이다.

최근에는 Transformer 기반 Self-attention 구조가 도입

되어 새로운 패러다임을 형성하고 있다. SegFormer,[18] InternImage[19] 등은 Self-attention을 통해 장거리 의존성과 다중 스케일 특징 간의 상호작용을 효과적으로 모델링하며, 기존 CNN 기반 모델 대비 정교한 구조 표현이 가능하다. 이러한 트랜스포머 기반 모델은 현재 의미론적 분할 분야에서 최첨단(State-of-the-art) 성능을 보여주고 있으며, 향후 해당 분야의 발전을 주도할 것으로 기대된다.

3. 서브-모달리티 기반 웨이블릿 변환을 이용한 심층 신경망

그림 1은 제안하는 네트워크의 전체 구조를 나타낸다. 본 모델은 크게 인코더 부분과 디코더 부분으로 구성된 인코더-디코더 아키텍처이며, 효율적인 처리를 위해 ERFNet을 기반으로 설계되었다. ERFNet은 실시간 의미론적 분할을 위해 고안된 경량화된 Residual 구조의 네트워크로, 인코더 단계에서 다단계의 특징지도를 추출하고 디코더 단계에서 해상도를 복원한다.

제안된 방법은 서브-모달리티 주의 (Sub-modal attention) 기반으로 ERFNet의 인코더를 변환하게 된다. 웨이블릿 변환 (Wavelet transform)을 이용하여 저주파 영역 신호와 고주파 영역 신호의 두 개의 서브-모달리티로 나누어서 사용하며, 얻어진 두 종류의 특징지도는 몇 단계의 계층을 거쳐서 상호-주의 모듈들을 통해 융합된다.

3.1 웨이블릿 변환(Wavelet transform)

웨이블릿 변환[20](Wavelet Transform, WT)은 시간 또는 공간 도메인에서의 다중 해상도 분석(Multi-resolution analysis)을 가능하게 하는 강력한 수학적 도구로, 주파수 영역을 분리하여 신호를 효과적으로 분석할 수 있다. 이 변환은 입력 신호를 저주파 성분(전역적인 정보)과 고주파 성분(세부적인 정보)으로 구성된 여러 개의 부대 역(Subband)으로 분해함으로써, 다양한 스케일에서의 특징 추출이 가능하다. 이러한 특성 덕분에 웨이블릿 변환은 에지 검출(Edge detection), 영상 압축(Image compression) 등 다양한 영상 처리 분야에서 활용되고 있다.

변환 과정에서 사용되는 LF(Low-pass Filter)와 HF (High-pass Filter)는 각각 저역 및 고역 통과 필터를 의미하며, 각 필터링은 합성곱(Convolution) 연산을 통해 수행된다. 구체적으로, 입력 영상 L에 대해 먼저 행 방향으로 LF 및 HF 필터링을 적용하고, 이어서 열 방향으로 동일한 필터링을 수행한다. 각 필터링 단계 이후에는 2배로 다운샘플링을 진행하여 주파수 성분을 분리한다.

이 과정을 통해 총 네 개의 부대 역 성분이 생성되며, 각 성분은 다음과 같은 의미가 있다. LL은 수평 및 수직 방향 모두 저역 필터를 적용한 결과로, 영상의 전역적 구조나 배경 정보를 포함한다. LH는 수평 방향으로는 저역, 수직 방향으로는 고역 필터를 적용한 결과로, 수평 방향의 세부 정보를 나타낸다.

HL은 수평 방향으로는 고역, 수직 방향으로는 저역 필터를 적용한 결과로, 수직 방향의 세부 정보를 포함한다. 마지막으로 HH는 고역 필터를 양방향 모두에 적용한 결과로, 대각선 방향의 고주파 성분을 포함한다. 또한, 웨이블릿 변환은 다단(Multi-level) 구조로 확장할 수 있다. 이를 위해 저주파 부대 역 성분인 LL에 대해 동일한 웨이블릿 분해 과정을 반복 적용함으로써, 더 낮은 해상도에서의 전역적 특징과 점점 더 세밀한 계층적 정보를 추출할 수 있다. 이러한 다중 스케일 표현은 영상 내의 다양한 공간 주파수 성분을 효과적으로 포착할 수 있어, 더욱 정교한 특징 표현이 요구되는 의미론적 분할, 복원, 압축 등 고차원 비전 과제에서 유용하게 활용된다.

다만, 본 논문에서는 웨이블릿 변환에 사용되는 필터로서 구조가 단순하고 계산량이 적은 Haar wavelet 필터뱅크를 채택하였고, 웨이블릿 변환은 한 번만 적용하였다. 이는 기본적인 주파수 성분 분해를 효율적으로 수행할 수 있어 다양한 비전 응용에 적합하다. 다단 구조를 통해 주파수 영역을 더 다양하게 나눌 수 있겠지만, 결과로 나오는 성분들의 해상도가 제각각으로 나타나기 때문에, 네트워크 구성의 직관성이 떨어지며 결합방식 또한 현재 방식으로 진행할 수 없기에 고려하지 않았다.

3.2 서브-모달리티 주의 융합 네트워크 (Sub-modal Attention Fusion Network)

서브-모달리티 주의 모듈은 고주파 성분과 저주파 성분 특징지도 간의 상호 보완성에 주목한다. 이는 RGB 입력 영상을 사용하여 전반적인 의미정보를 추출하는 일반적 단일-모달리티(Uni-modal) 기반 자기 -관심(Self-attention)에서 나아가, 성질이 다른 성분인 고주파 및 저주파 성분으로 나누어서 다중-모달리티(Multi-modal) 개념을 활용하여 이들 간의 관계성을 모델링하였다.

제안하는 알고리즘의 전체 구조는 Fig. 1에서 확인할 수 있다. 우선, 웨이블릿 변환을 이용해 얻은 저주파 성분(Low Frequency Feature Map, LFFM)과 고주파 성분(High Frequency Feature Map, HFFM)을 각각 LFFM={LL}과 HFFM={LH, HL, HH}으로 분류한다. 두 성분은 각각 저주파 성분 특징 추출 계층(Low Frequency Feature Extraction Layer, LF FEL)과 고주파 성분 특징 추출 계층(High Frequency Feature Extraction Layer, HF FEL)을 지나 각 성분에 해당하는 특징지도를 얻는다.

각 주파수 성분에 맞는 처리가 된 특징지도들은 Fig. 2의 서브-모달리티 주의 융합 모듈 (Sub-Modal Attention

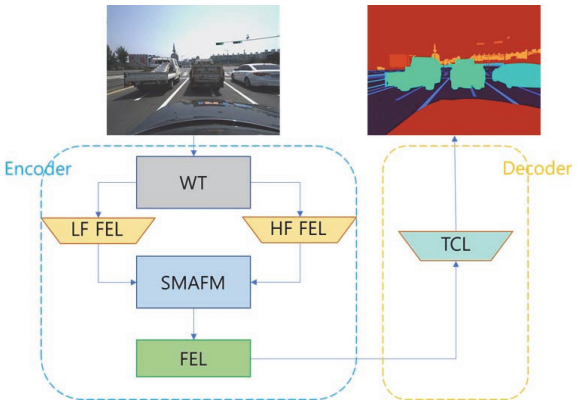

Fig. 1 Proposed network architecture

Fusion Module, SMAFM)을 거쳐서 결합하게 된다. SMAFM은 두 모달리티를 효과적으로 결합하기 위해 제안되었으며, 각 모달리티의 정보가 서로를 동적으로 강화해 줄 수 있도록 주의(Attention) 메커니즘을 기반으로 설계되었다.

우선 두 입력 특징지도에 대해 먼저 공간적 중요 영역을 강조하기 위해 공간적 주의(Spatial attention) 모듈을 각각 적용한다. 이는 CBAM의 공간적 주의 설계를 기반으로 하며, 입력 특징지도 $x \in R^{B\times C\times H\times W}$에 대해 다음과 같이 수행된다:

$$SA(x)=\sigma(Conv_{5\times 5}([AvgPool(x);MaxPool(x)]))\odot x \tag{1}$$

여기서 σ는 시그모이드 함수, [;]는 채널 축으로의 연결(Concatenation)을 나타내며, $\odot$는 요소별 곱을 나타낸다. 평균 풀링과 최대 풀링을 결합함으로써, 보다 넓은 공간적 패턴과 국소적인 주목 대상을 동시에 반영한다.

공간적 주의로 보정된 특징지도 x'_L,x'_H에 대해, 각 모달리티는 상대모달리티의 전역정보(Global Average Pooling, GAP) 기반하여 채널 요약정보를 만든다:

$$a_L=\sigma(W_L^{(2)}\cdot ReLU(W_L^{(1)}\cdot GAP(x'_H))) \tag{2}$$

$$a_H=\sigma(W_H^{(2)}\cdot ReLU(W_H^{(1)}\cdot GAP(x'_L))) \tag{3}$$

여기서 $W_*^{(1)}\in R^{C\to C/r}$, $W_*^{(2)}\in R^{C/r\to C}$는 1×1 컨볼루션에 해당하며 r은 채널 축에 대한 축소비율이다 (제안된 방법에서는 r=16).

이를 기반으로 자신의 채널별 중요도를 동적으로 재조정한다(Fig. 2):

$$x_L^a=x'_L\odot a_L \tag{4}$$

$$x_L^a=x'_L\odot a_L \tag{5}$$

이러한 구조는 모달리티 간의 단순 결합 대비, 더 세밀한 의미 기반 융합을 가능하게 한다.

기반이 되는 ERFNet의 인코더 부분의 초기 몇 개의 계층이 상술한 계층들로 대체가 되며, 이후의 특징 추출 계층(Feature Extraction Layer, FEL) 및 디코더에 해당하는 전치 합성곱 계층(Transposed Convolution Layer, TCL) 구성은 ERFNet과 동일하다.

Table 1에서는 각각의 계층들에 대해 입출력 특징지도의 크기를 고려해 가며, 구체적인 네트워크 구조를 성명한다. 입력 데이터는 실험에서 사용된 Cityscapes[21] 데이터 세트의 입력 영상을 절반으로 리사이즈 한 크기인 1024 × 512 (너비 × 높이) 가 기준이 된다.

입력 영상이 들어오면 RGB 각각의 채널에 대해, Haar 필터뱅크를 이용하여 신호를 고주파 성분과 저주파 성분으로 나눈다. 고주파 성분인 HFFM과 저주파 성분인 LFFM은 원래 채널 수의 3배수와 1배수를 가지게 되므로 각각의 채널 수는 9와 3이 된다. 그리고 각각의 성분의 특징지도는 LF FEL과 HF FEL을 통과하며 다시 절반의 크기가 되는데, 두 계층 모두 1 × 1 커널 크기를 가진 합성곱을 사용하여 정보를 정제해주고, 스트라이드 (stride)

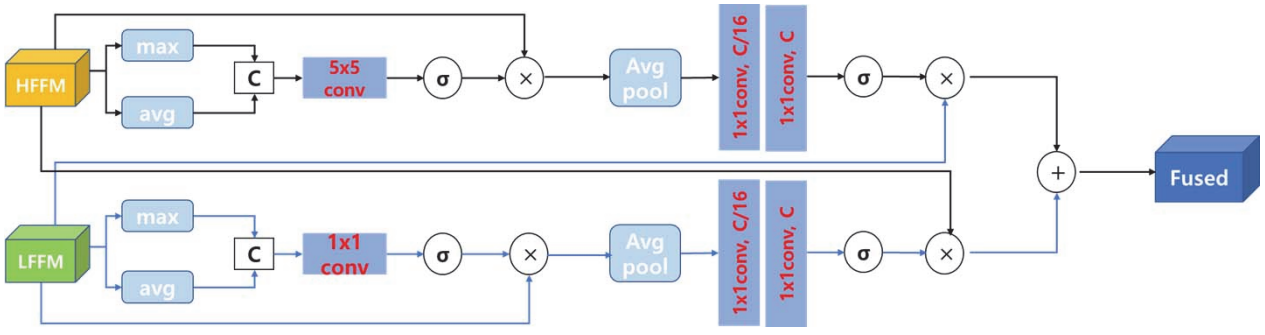

Fig. 2 Sub-modal attention fusion module

Table 1 Detailed architecture of the proposed semantic segmentation

		Input	Operators	Output
	WT	1024 × 512 × 3	Wavelet transform	512 × 256 × 3
				512 × 256 × 9
Encoder	LF FEL	512 × 256 × 3	1×1 conv	512 × 256 × 16
		512 × 256 × 16	Downsampler block	256 × 128 × 64
	HF FEL	512 × 256 × 9	1×1 conv	512 × 256 × 16
		512 × 256 × 16	Downsampler block	256 × 128 × 64
	SMAFM	256 × 128 × 64, 256 × 128 × 64	Sub-Modal Attention Fusion	256 × 128 × 64
	FEL	256 × 128 × 64	3×3 conv (5 times)	256 × 128 × 64
		256 × 128 × 64	Downsampler block	128 × 64 × 128
		128 × 64 × 128	3×3 conv (8 times)	128 × 64 × 128
Decoder	TCL	128 × 64 × 128	T-convolution	256 × 128 × 64
		256 × 128 × 64	3×3 conv (2 times)	256 × 128 × 64
		256 × 128 × 64	T-convolution	512 × 256 × 16
		512 × 256 × 16	3×3 conv (2 times)	512 × 256 × 16
		512 × 256 × 16	T-convolution	1024 × 512 × N_C

가 2인 3 × 3 합성곱을 통과시켜서 해당 기능을 수행한다.

SMAFM의 경우에는 LFFM과 HFFM을 합해주게 되는데, 두 입력의 특징 크기가 256 × 128 × 64로 일치한 상태에서 각 입력에 대해 서로의 중요도를 재조정해주게 된 상태로, 두 신호를 더해준 다음 ReLU를 통과시켜서 다음 과정으로 넘겨주게 된다.

이후의 과정은, 원형이 되는 네트워크 모델인 ERFNet과 동일하다. Dilation이 없는 3 × 3 Convolution에 연속해서 Dilation이 있는 3 × 3 Convolution을 수행하는 3 × 3 conv를 5회 (Dilation은 1로 동일) 통과시키고, 추가적 Downsampler block을 통해 특징지도 크기를 한 번 더 축소한 뒤에 다시 3 × 3 conv를 8회 통과시킴으로써 (Dilation은 차례대로 2, 4, 8, 16, 2, 4, 8, 16), 정보를 압축하는 인코더에 해당하는 역할을 끝내게 되며 128 × 64 × 128 크기의 특징지도를 얻게 된다. 이렇게 인코딩된 정보를 디코더를 통해 최종 결과인 라벨 지도를 얻게 되는데, Stride가 2인 Transposed convolution(T-convolution)과, 이어지는 3 × 3 conv(Dilation 1) 들로 구성된 Transposed Convolution Layer(TCL)을 통과시켜서 N_c개 클래스에 대한 확률이 모든 화소에 담겨있는 최종 결과를 생성한다.

제안된 방법은 원형이 되는 네트워크에서 변형하였기에, 네트워크와 관련된 파라미터나 구조는 ERFNet 설정과 같게 사용할 수 있었으며 추가 적으로 고려 되어야 할 사항들은 다음과 같았다. 제안된 방식은 LF 성분과 HF 성분을 2번의 FEL을 각각 거친 다음 결합하였는데, 이는 실험적으로 결정되었으며, 1번 혹은 3번 거치는 구조는 성능이 소폭 하락하는 것을 확인하였다. SMAFM의 채널 축소비율인 r 또한 실험적으로 16이 최적으로 채택되었다.

4. 실 험

4.1 데이터 세트

본 논문에서는 차량에서 촬영된 영상들로 구성되어 자율주행에 활용가능한 Cityscapes[21], KITTI-360[22] 데이터 세트와 더불어 국내 도로 주행데이터로 구성된 Multi Camera Semantic Segmentation 데이터 세트(MCSS, https://nanum.etri.re.kr/에서 제공)를 이용해 제안된 모델의 성능을 검증하였다.

Cityscapes 데이터 세트는 독일 50개 도시에서 촬영된 도심 도로 주행 영상 위주로 구성되어 있으며 5,000장의 2048 × 1024 해상도의 영상들이 19개의 주요 클래스로 주석 파일이 있으며 이를 이용해 성능평가 기준을 적용한다.

KITTI-360은 기존 KITTI 데이터 세트를 확장하여 만들었으며, 도시/도심/교외 등 다양한 환경에서 영상뿐만 아니라, 라이다 포인트 클라우드 및 스테레오 카메라의 정보들이 포함되어 있다. 영상은 1408 × 376 크기로 총 9개의 시퀀스며, 클래스 정보는 Cityscapes와 같으나, 데이터상으로 거의 존재하지 않는 Bus와 Train은 실험에서 제외하였다. 본 논문에서는 스테레오 영상이 페어로 존재하며, LiDAR 정보도 함께 존재하는 61,186장의 영상을 데이터 세트로 활용하였다.

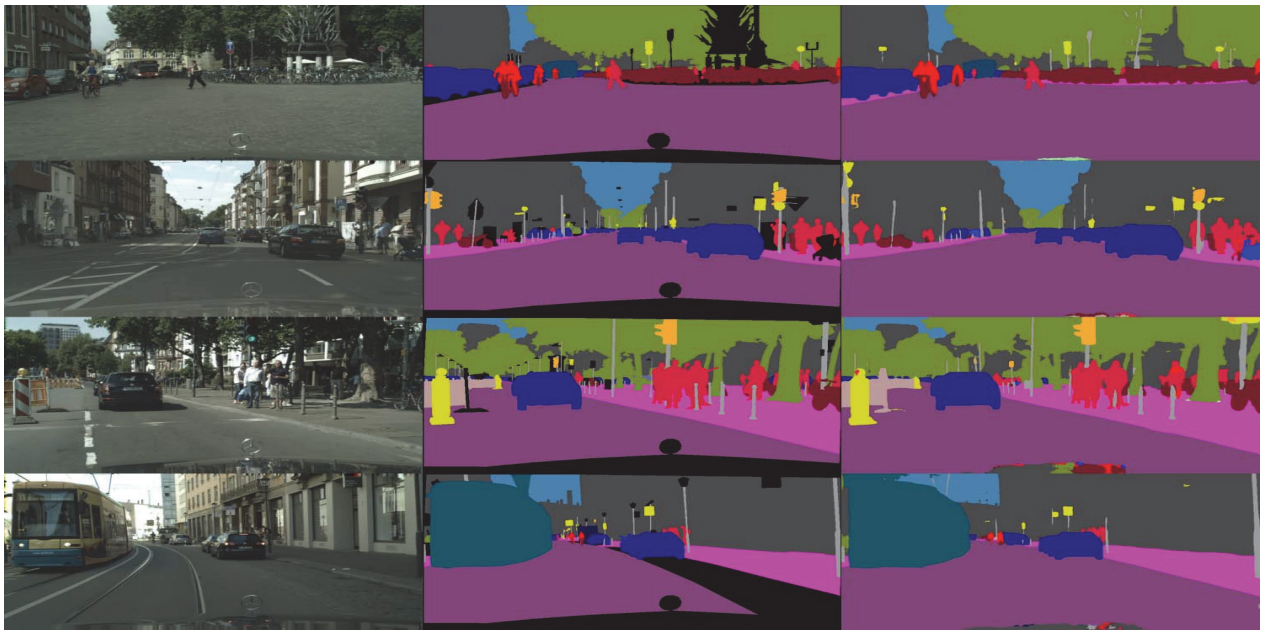

Fig. 3 Qualitative results on cityscapes dataset

MCSS 데이터 세트는 국내의 여러 장소를 주행하며 촬영한 692장(학습: 512, 검증 182장, 해상도 2048 × 1536)의 데이터를 이름처럼 다중 카메라인 Stereo 데이터로 제공하고 있고, 기준 되는 카메라에 해당하는 참값 정보를 함께 가지고 있다. 원래 제공되는 참값 정보로는 62종의 다양한 클래스를 기준으로 분류하고 있으나, 한정된 데이터 숫자에 비해 많은 클래스 숫자와 실험의 편의를 고려하여 이들을 19개의 그룹으로 통폐합하여 클래스로 삼아 실험을 진행하였다. 통합하여 사용된 클래스는 도로, 보도, 로드 마크, 차선, 연석, 방벽, 차, 트럭, 버스, 2륜 탈 것, 기타 차량, 보행자, 라이더, 라바콘, 건물, 교통표지판, 신호등, 수직의 물체들, 기타 영역으로 나누어서 사용하였다.

4.2 네트워크 학습

네트워크 학습환경은 우분투에서 Pytorch를 사용하였으며, 개발용 PC 사양은 Intel i9 10980XE, NVIDIA RTX 3090, 128GB RAM이다.

Cityscapes와 KITTI-360을 학습 함에 있어 공통으로 SGD 최적화 함수를 사용하였으며, 초기 학습률 (Learning

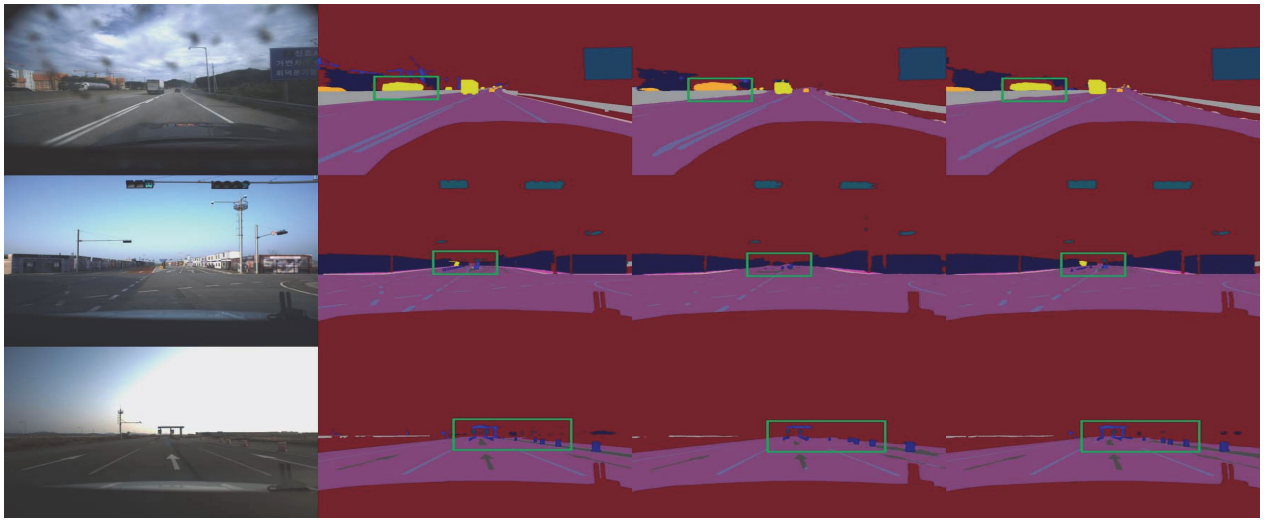

Fig. 4 Qualitative results on MCSS dataset

rate)는 각각 5e-3과 1e-2로 설정하였다. 학습에 사용된 손실 함수는 Cross-entropy이며, 500 epoch 동안 학습을 수행하였다.

Cityscapes의 경우 학습용 데이터 2,975장과 검증용 데이터 500장을 절반의 크기인 1024 × 512로 학습 및 실험 결과를 관찰하였다. KITTI-360의 경우에는 9개의 영상 시퀀스 중 시퀀스 번호 0, 2, 3, 4, 5, 6을 학습에 사용하고 (총 45,108장), 시퀀스 번호 7, 9, 10을 검증에 사용하였는데 (총 16,060장), 영상 시퀀스의 특성상 중복되는 장면이 많기에 10개 단위로 샘플링하여 학습과 검증에 사용되는 시간을 줄였다. MCSS 데이터 세트의 경우는 학습하기에 영상의 숫자가 많은 데이터는 아니라고 생각되어 학습 초깃값으로 Cityscapes에서 학습한 결과물을 주고, 초기 학습률은 1.5e-2로 하였으며 나머지는 Cityscapes의 경우와 같게 설정하였다. 영상의 크기는 절반인 1024 × 768로 하여 학습 및 검증을 수행하였다.

4.3 실험 결과 및 평가

본 논문에서는, 제안한 모델의 성능을 평가하기 위해 mIoU(mean Intersection over Union)와, 모델의 연산속도를 측정하였다. IoU는 클래스별로 사용되는 정확도로, 모델이 예측한 영역과 정답 영역의 겹치는 부분과 전체 영역 간의 비율을 측정하며, mIoU는 앞서 구한 IoU들의 클래스들에 대한 평균값이다.

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$mIoU = \frac{1}{N_c}\sum_{c=1}^{N_c} IoU_c \tag{7}$$

여기서, 예측과 정답 영역이 겹치는 영역은 TP(True Positive)이고, 전체 영역은 TP와 FP(False Positive) 및 FN (False Negative)의 합 영역이다.

또한, 네트워크의 성능 대비 효율을 알아보기 위해, 실제 네트워크에 입력 영상이 들어간 후부터 모델을 통과하기까지 걸리는 시간을 관찰하였으며, 이는 개발용 PC에서 같은 크기의 입력을 100회 통과시키는데 걸리는 평균 연산시간을 이용하였다. 실험에 사용된 비교 대상들 또한 같은 환경에서 실험을 진행하였다.

제안된 서브-모달리티 주의 기반 네트워크의 발전성을 확인하기 위해, 모델의 기본이 되는 ERFNet과, 제안된 방법처럼 웨이블릿 변환을 통해 신호를 분리한 뒤 단순 결합방식을 수행한 방법인 An과 Choi[13]의 방식과 결과를 비교하였다.

Tables 2 ~ 4는 각각 Cityscapes와 KITTI-360, MCSS 데이터 세트에서 실험한 결과를 정량적으로 나타내고 있다. 실험에서 모두 입력단 계층을 효과적으로 바꿔줌으로써, 연산시간 단축과 동시에 성능개선이 이루어지는 것을 확인할 수 있다. Cityscapes의 경우, 네트워크가 영상 1장 처리에 걸리는 시간이 8.3 ms에서 7.2 ms로 줄어들고, 성능은 70.85 %에서 74.72 %로 3.87 % 상승하였다. 속도 상승에는 다음과 같은 요인들이 있다. 첫 번째로, 웨이블릿 변환을 통해 해상도가 절반으로 줄어든 상태에서 합성곱 작업을 수행하고, 합성곱의 크기 또한 3 × 3에서 1 × 1로 줄어들어서 계산량이 줄어들게 된다. 웨이블릿 변환 자체의 연산 비용은 그에 비해 비교적 작다. 두 번째로는, 경로가 두 개로 나누어져서 연산 자체는 두 번씩 해야 하지만 GPU에서 두 경로를 병렬적으로 처리할 수 있기에 더 빠르게 실행될 수 있다. 기존 단순 덧셈식의 결합방식에 비해 SMAFM 만큼 연산시간이 늘어나지만, 기존에도 실시간성이 충분한 7.0 ms에서 0.2 ms인 3 % 정도의 시간을 추가로 소요해서 1.7 %가량의 추가 성능 향상을 이루었으므로 합리적인 비용이라 할 수 있다. Fig. 3은 제안된 방법으로 Cityscapes 데이터 세트에서의 정성적 실험 결과이다. 각 그림은 입력 영상(왼쪽), 참값(중간), 제안된 네트워크로 수행한 의미론적 분할의 결과로 구성되어 있다. 검출 결과에서 각각의 색은 레이블 클래스를 나타내며, 영상들은 큰 구조를 나타내는 건물, 도로 외에도 다양한 물체들을 볼 수 있게 선정하였으

Table 2 Quantitative results with cityscapes dataset

	mIoU	Computation time
ERFNet	0.7085	8.3 ms
Method in 13)	0.7305	7.0 ms
Proposed method	0.7472	7.2 ms

Table 3 Quantitative results with KITTI-360 dataset

	mIoU	Computation time
ERFNet	0.5442	8.5 ms
Method in 13)	0.5663	7.1 ms
Proposed method	0.5924	7.3 ms

Table 4 Quantitative results with MCSS dataset

	mIoU	Computation time
ERFNet	0.5095	13.15 ms
Method in 13)	0.5709	10.68 ms
Proposed method	0.5816	10.98 ms

며, 큰 구조물뿐만 아니라, 사람이나 오토바이, 폴대나 표지판 등과 같은 작은 물체 영역도 잘 뽑아내고 있음을 확인할 수 있다. 이 그림을 통해, 제안된 방법을 사용하면 다양한 상황에 대해 의미론적 분할을 안정적으로 수행한다는 것을 확인할 수 있다.

KITTI-360의 경우에도 원형이 되는 ERFNet에 비해서는 영상 1장 처리에 걸리는 시간은 8.5 ms에서 7.1 ms로 줄어들고, 성능은 54.52 %에서 59.24 %로 4.82 % 상승하였다. 단순 덧셈 방식의 결합방식에 비해서도 0.2 ms를 추가로 소모하여 2.61 % 성능 향상을 달성하였다.

MCSS 데이터 세트에서도 앞의 데이터들과 같은 경향성이 나타나는 것을 확인할 수 있으며, Fig. 4는 제안된 방법으로 MCSS 데이터 세트에서 실험한 결과를 정성적으로 나타내고 있다. 각 그림은 왼쪽부터 입력 영상, 참값, ERFNet의 결과, 제안된 방법의 결과를 나타낸다. 전체적으로 세그멘테이션 결과가 세밀하게 나타나는 것을 볼 수 있다. 특히 초록 상자 안은 세밀한 결과뿐 아니라, ERFNet에서는 찾지 못하는 것을 제안된 방법으로 찾는 것을 확인할 수 있다.

5. 결 론

본 논문에서는 실시간 용도로 사용되기 적합한 모델에서 입력 영상을 주파수 특징에 따라 나누어서 서브-모달리티를 구축하고, 서브-모달리티 주의 기반의 특징지도 결합 방법으로 구성된 의미론적 영상 분할 인공신경망 구조를 제안하였다. 계층 구조를 깊게 복잡하게 만드는 것 외에, 입력 데이터를 분석, 분해하여 효율적으로 사용하는 방식을 보였으며, 다양한 데이터 세트에서 연산시간 감소와 성능 향상이라는 일관된 경향성을 실험적으로 입증하였다. 제안된 방식에서는 입력 영상을 웨이블릿 변환을 통해 주파수 성분으로 분해하였지만, 다양한 색 공간들이나 추가적인 성분들로 영상을 분해하여 실험할 수 있을 것이다.

후 기

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

본 논문은 한국전자통신연구원에서 ETRI AI 나눔을 통해 공개한 [자율주행]Multi Camera Semantic Segmentation 학습 데이터 학습 데이터 셋을 사용함.

References

1) O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp.234–241, 2015.
2) J. Long, E. Shelhamer and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3431–3440, 2015.
3) V. Badrinarayanan, A. Kendall and R. Cipolla, “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, 2017.
4) L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, “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, 2017.
5) H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, “Pyramid Scene Parsing Network,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2881–2890, 2017.
6) E. Romera, J. M. Alvarez, L. M. Bergasa and R. Arroyo, “ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation,” IEEE Transactions on Intelligent Transportation Systems, Vol.19, No.1, pp.263–272, 2017.
7) C. Yu, J. Wang, C. Peng, C. Gao, G. Yu and N. Sang, “BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation,” Proceedings of the European Conference on Computer Vision (ECCV), pp.325–341, 2018.
8) H. Pan, C. Song, W. Jiang, X. Lu and C. Li, “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, 2022.
9) M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo and X. Wei, “Rethinking BiSeNet for Real-Time Semantic Segmentation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.9716–9725, 2021.
10) G. Cho, C. Ahn and J. Gim, “Improving Nighttime Curb Segmentation with Domain Translation,”

Transactions of KSAE, Vol.32, No.8, pp.657–666, 2024.

11) Y. J. Chang and J. K. Suhr, “Single CNN-Based Simultaneous Parking Slot Detection and Segmentation Using AVM Images,” Transactions of KSAE, Vol.32, No.7, pp.559–568, 2024.
12) Y. Lee, J. Jeon, J. Yoon and K. Song, “3D Reconstruction of 2D Semantic Segmentation Using Sensor Fusion Based Trans-UNet,” Transactions of KSAE, Vol.31, No.4, pp.255–263, 2023.
13) T.-H. An and J. D. Choi, “Efficient Semantic Segmentation Using Wavelet-Transform,” Journal of the Korea Institute of Intelligent Transport Systems, Vol.23, No.5, pp.248–260, 2024.
14) J. Kang, S. J. Han, N. Kim and K. W. Min, “ETLi: Efficiently Annotated Traffic LiDAR Dataset Using Incremental and Suggestive Annotation,” ETRI Journal, Vol.43, No.4, pp.630–639, 2021.
15) J. Hu, L. Shen and G. Sun, “Squeeze-and-Excitation Networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.7132–7141, 2018.
16) Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, “ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.11534–11542, 2020.
17) S. Woo, J. Park, J. Y. Lee and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” Proceedings of the European Conference on Computer Vision (ECCV), pp.3–19, 2018.
18) E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” Advances in Neural Information Processing Systems (NeurIPS), Vol.34, pp.12077–12090, 2021.
19) W. Wang, J. Xie, Z. Yu, J. Alvarez, A. Anandkumar and P. Luo, “InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.14408–14419, 2023.
20) M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice Hall, 1995.
21) M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson and B. Schiele, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3213–3223, 2016.
22) Y. Liao, J. Xie and A. Geiger, “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, 2022.