https://doi.org/10.7840/kics.2025.50.1.22

실시간 영상 분할을 위한 TBNs 구조 기반 이중 방향 다중센서 융합 기법

신 윤 식[◆], 손 용 호[*], 박 정 희[**], 이 채 현[**], 김 양 곤[**], 최 준 원[°]

TBNs-Based Bidirectional Fusion Method For Real-Time Multi-Modal Segmentation

Yunsik Shin[◆], Yongho Son[*], Junghee Park[**], Chaehyun Lee[**], Yanggon Kim[**], Jun Won Choi[°]

요 약

자율주행 및 로보틱스 분야에서 주변 환경과 객체의 형태를 인지하는 영상 분할 기술은 필수적인 요소로 자리 잡고 있다. 카메라를 사용한 영상 분할 기술은 다양한 색 정보를 통해 영상 분할을 수행하지만 환경에 대한 공간적 정보를 활용하지 못하는 한계점이 존재한다. 이를 위해 깊이 영상 또는 라이다 등의 센서 데이터를 융합하여 보다 신뢰성 높은 성능을 달성하기 위한 연구가 진행되었다. 하지만 대부분의 연구는 높은 성능 달성을 위해 복잡한 구조의 융합 기법을 사용하며 제한된 실시간 동작 환경에서 구현이 불가능한 문제점을 내포한다. 본 연구는 카메라 영상과 깊이 영상의 특성을 활용하여 실시간 동작이 가능한 효율적 융합 기법을 제안한다. 제안하는 이중 방향 융합 모델은 카메라 모델보다 1.27 mIoU 성능 향상을 달성함과 동시에 16.32 FPS 동작속도를 보이며 실시간 동작이 가능함을 보였다.

키워드 : 영상 분할, 멀티모달 융합, 스테레오 카메라 깊이 영상 융합
Key Words : 2D Semantic Segmentation, Multi-modal Fusion, Depth map Fusion

ABSTRACT

In the field of autonomous driving and robotics, image segmentation that perceives the shape of the surrounding environment and objects has become an essential part. While image segmentation using cameras performs segmentation through various color information, it has limitations in utilizing spatial information of the environment. To address this, research has been conducted to achieve higher reliability by fusing sensor data such as depth images or LiDAR. However, most studies involve complex fusion techniques to achieve high performance, which inherently poses problems that are not feasible to implement in limited real-time operational environments. This study proposes an efficient fusion technique that leverages the characteristics of camera and depth images to enable real-time operation. The proposed bidirectional fusion model achieves a performance improvement of 1.27 mIoU over the camera model while showing an operating speed of 16.32 FPS, making real-time operation possible.

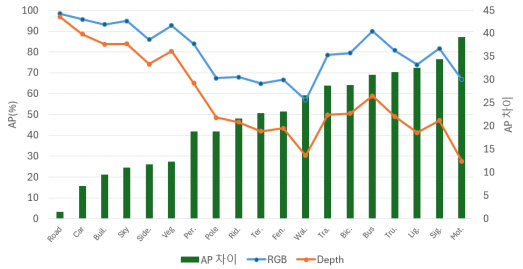
※ 이 논문은 2022년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(No. KRIT-CT-22-011-00, LiDAR-Camera Image Fusion).
◆ First Author : Hanyang University Department of Electrical Engineering, ysshin@spa.hanyang.ac.kr, 학생회원
° Corresponding Author : Seoul National University Department of Electrical and Computer Engineering, junwchoi@snu.ac.kr, 종신회원
* Hanyang University, yhson@hanyang.ac.kr,
** LIGNex1, junghee.park@lignex1.com; chaehyun.lee@lignex1.com; yanggon.kim@lignex1.com
논문번호 : 202402-033-A-RN, Received February 21, 2024; Revised August 13, 2024; Accepted September 9, 2024

Ⅰ. 서 론

영상 분할은 객체 검출과 함께 자율 주행, 로보틱스 분야에서 주변 환경을 인지하기 위한 필수적인 역할을 수행한다. 보다 신뢰성 있는 검출 결과를 위해 카메라, 깊이 영상, 라이다 등의 센서를 동시에 입력 받아 색 정보와 공간적 정보를 함께 활용하는 멀티모달 영상 분할 모델들이 연구되었으며, 카메라 영상만을 활용할 때보다 높은 성능을 보이는 연구 결과들을 확인할 수 있다[9,13,15]. 하지만 높은 성능을 위해 복잡한 모델 구조, 융합 모듈들이 연구됨에 따라 실시간 동작 조건에 부합하지 않는 연구들이 대다수를 차지하고 있으며, 실제 자율 주행 환경에서 연구 성과를 활용하기 힘든 문제가 존재한다. 요구량이 증가하는 연구 흐름은 지속되고 있으며, 영상 분야에서의 초기 모델인 BiseNet[14]은 1.4백만개의 파라미터 수를 가졌지만 최근 InternImage[12]모델의 경우 1억개의 파라미터 수를 갖을 정도로 모델의 크기가 가파르게 증가하는 추세를 보인다.

초기 멀티모달 영상 분할 모델의 경우 깊이 인코더에서 카메라 인코더 방향으로 특징 지도를 융합하는 방식을 사용하였다. 이러한 구조는 카메라 모델의 성능이 비교적 높은 특성을 고려한 것이며, 깊이 정보를 보완적 데이터로 활용하고자 하는 모델 설계 철학에 기인한다. 이러한 단방향 모델 설계 방식은 단순한 구조를 사용하여 낮은 연산 복잡성을 보이기 때문에 실시간 동작에 적합하지만 다소 낮은 성능을 보인다는 단점을 가진다. 최근 연구는 초기 연구보다 복잡한 구조인 양방향 융합 구조를 사용하는 흐름을 보인다. 양방향 융합 구조는 카메라, 깊이 영상 데이터를 각각 입력 받은 후 추출된 특징 지도를 융합 모듈에 입력으로 사용하여 융합된 특징 지도를 얻는다. 높은 성능을 위해 더욱 복잡한 융합 모듈 구조가 사용되고 있으며, 많은 메모리와 연산량, 연산 시간을 요구하므로 실제 환경에서 동작하기 힘든 문제점이 존재한다.

그래프 1은 카메라, 깊이 영상 모델의 영상 분할 성능 차이를 표현한다. 카메라 모델의 경우 모든 클래스에 대해 비교적 높은 성능을 보이는 모습을 확인할 수 있다. 깊이 영상 모델의 경우 도로, 건물, 하늘, 인도, 식생 등의 넓은 영역의 클래스에 대해서는 카메라 모델 대비 성능 차이가 적은 반면, 오토바이, 신호등, 교통 표지판 등의 작은 객체에 대해서는 성능 하락이 많은 모습을 보인다. 즉, 깊이 영상 모델이 넓은 영역에서 강건한 모습을 보이며, 이러한 특성은 그림 1처럼 시각화 된다. 그림 1은 순서대로 정답 레이블, 카메라 모델의 출력, 깊이 영상 모델의 출력을 의미한다. 좌우측에 수풀이



Graph 1. Performance Comparison of camera, depth segmentation model

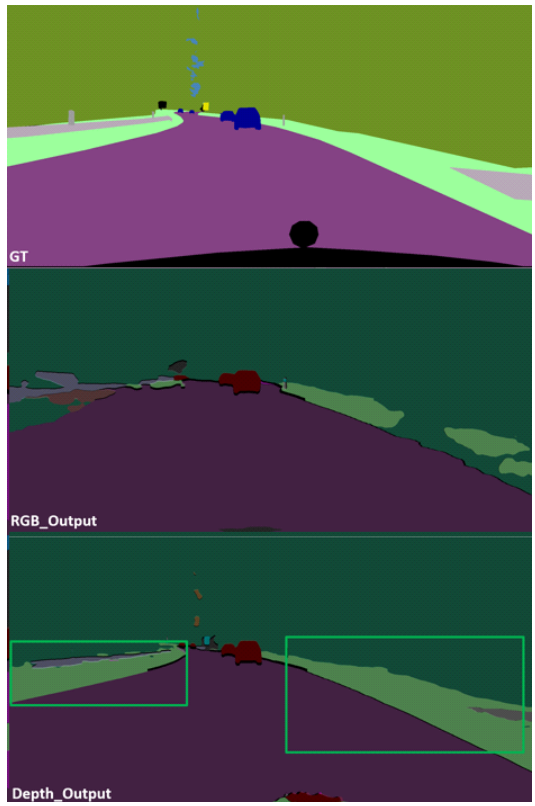


Fig. 1. Segmentation Results from camera, depth segmentation model

존재하지만 카메라 모델의 경우 해당 영역을 정확히 분할하지 못하였다. 특히 우측의 경우 수풀 영역을 색 질감이 비슷한 식생 영역으로 오판하는 모습을 확인 가능하다. 깊이 영상 모델의 경우 보다 신뢰성 있는 결과를 출력하는 모습이 확인 가능하다.

논문에서는 높은 영상 분할 성능과 실시간성을 동시에 얻을 수 있는 융합 기법을 제안한다. 제안하는 이중 방향 융합 기법은 위의 그래프 1, 그림 1에서 확인 가능한 카메라 영상 및 깊이 영상 데이터의 특성을 활용함으

로써 단방향 융합 기법의 낮은 연산량 장점과 및 양방향 융합 기법의 높은 성능 장점을 획득한다 이를 위해 우선 2절에서 카메라, 깊이 영상 기반 영상 분할 연구들에 대해 소개한다. 이후 3절 이중 방향 융합 기법에서 다중 브랜치 네트워크의 개념을 소개하고, 본 논문이 제안하는 이중 방향 융합 기법에 대해 소개한다. 마지막으로 실험 결과와 분석을 통해 제안하는 기법의 유효성에 대해 정리한다. 제안된 기법은 카메라 모델 대비 1.27 mIoU 성능 향상을 보이며, 16.32 FPS의 연산 속도를 보였다. 또한 모델의 성능 결과를 분석하여 제안하는 모델이 비교 대상 모델들의 문제점을 어떻게 해결하였는지 보고한다.

Ⅱ. 영상 분할 기술 동향

2.1 RGB 영상분할

연산곱 네트워크 (Convolutional Network)[11]는 컴퓨터 비전 테스크를 수행하기 위한 모델의 기본 구조로 활용된다. 연산곱 네트워크는 풀링 레이어(pooling layer)를 통해 특징 지도의 크기를 점차 줄이며, 입력 영상의 함축적이고 중요한 정보를 추출하여 연상량을 줄이는 효과를 가져온다. 영상 분할 모델은 인코더와 디코더로 구성된다. 인코더는 카메라 영상 입력의 특징 지도를 출력하며, 디코더는 특징 지도를 입력으로 받아 픽셀별 클래스 예측을 수행한다. PSPNet[1]은 피라미드 풀링 (pyramid pooling)을 사용한 후 ResNet[2]기반 연산곱 네트워크를 이용하여 영상 분할을 위한 전역적 문맥 정보를 담은 특징지도를 추출하였다. 이러한 구조를 통해 전역적 정보를 사용할 수 있으며, 보다 높은 성능을 보이는 영상 분할 모델을 설계할 수 있다. 최근 제안된 트랜스포머(transformer)[3] 기반 구조는 영상 분할에서 보다 높은 성능을 보이고 있으며, 입력 영상의 전역적 정보를 고려하여 중요한 정보만을 포함한 특징지도를 추출한다. Segformer[4]는 위치 인코딩 정보가 필요하지 않은 계층 구조의 트랜스포머 디코더 네트워크를 제안하였으며, 높은 영상 분할 성능을 보이지만 실시간성 동작을 크게 고려하지 않고 설계된 단점이 있다.

영상 분할의 실시간 동작을 위해 설계된 모델로는 PIDNet[5]이 있다. 해당 모델은 세 가지 브랜치를 사용하며 각각 문맥적, 공간적, 윤곽 정보를 융합하여 사용하는 것이 특징이다. 해당 모델은 높은 영상 분할 정확도를 보이면서 실시간 동작성을 고려하였다.

2.2 RGB-D 영상 분할

RGB-D 영상 분할은 카메라, 깊이 영상, 라이다 등의 데이터를 융합하여 추출한 특징지도를 기반으로 영상 분할하는 것을 의미한다. 카메라 영상에는 깊이 정보가 존재하지 않으며, 깊이 영상 및 라이다 데이터에는 색 정보가 존재하지 않기 때문에 상호 보완적 관계를 활용하여 분할 성능을 높일 수 있다. 센서 융합을 수행하는 단계는 early-stage, middle-stage, late-stage 융합 방법으로 나눠진다. Early-stage 융합 방법으로는 ShapeConv[6]가 존재하며, 카메라 영상 특징지도와 깊이 영상 특징지도를 특징 지도의 채널 방향으로 합친 후 디코더로 전달한다. 해당 모델은 Shape-aware convolution 연산을 수행하여 객체의 형상 정보를 고려한 특징지도를 추출하는 것이 특징이다. 하지만 두 모달리티의 서로 다른 특징을 하나의 공유 네트워크에서 개별적으로 추출해내기 어렵다는 한계가 있다. Late-stage 융합 방법은 모델 아키텍처의 마지막 디코더 부분에서 융합이 이루어진다. Geometry-Aware Distillation[7]은 디코더 부분에서 깊이 영상의 기하 정보를 추출하여 융합하는 방식을 사용하였다. Middle-stage 융합방법은 세가지 융합 방법론들 중 가장 높은 성능을 보여주고 있는 방식이다. EMSANet[8]은 인코더와 디코더 네트워크에서 센서 융합을 수행하며, 실시간 동작을 위해 ResNet 네트워크 구조를 non-bottleneck 합성곱 연산으로 변경한 것이 특징이다. Middle-stage 융합 방법을 통해 높은 성능을 달성하였으며, 로봇 플랫폼에서도 실시간성을 보장한다. TokenFusion[9]은 트랜스포머 기반으로 middle-stage 융합을 수행한다. 인코더 네트워크에서 어텐션(attention) 구조를 사용하였으며, 각 영상에서 중요 정보를 추출한 후 특징 지도 융합을 수행한다.

2.3 RGB-D 융합 방향성

영상 분할을 위한 멀티모달 융합 기법은 단방향과 양방향 방식이 존재한다. 단방향 융합 기법은 서로 다른 두 개의 모달리티 영상의 특징지도를 한 방향으로만 융합하는 방식을 의미한다. 영상 분할의 경우, 카메라 영상 기반 모델이 깊이 영상 기반 모델보다 더 높은 성능을 보이므로 깊이 영상의 특징 지도를 카메라 영상의 특징 지도에 융합하여 사용하는 방법을 사용하고 있다. RFNet[16]의 경우 카메라 영상과 깊이 영상을 사용하지만 실시간 동작을 위해 단방향 융합 기법을 사용하였으며, 공간적 풀링(spatial pooling)을 사용하여 계산량을 더욱 낮춘 것이 특징이다. 해당 모델은 빠른 실시간성을 보여주지만 양방향 융합 기법 모델에 비해 낮은 성능을 가진다. 단방향 융합을 통해 깊이 영상의 문맥적(context) 정보를 전달하여 성능 향상을 보여주지만 깊이 영상 네트워크는 카메라 영상 정보와의 융합에 의한

성능 향상이 불가능한 단점이 존재한다. 양방향 융합 기법은 두 개의 다른 모달리티를 사용하는 네트워크 모두 융합하는 방법을 의미한다. 카메라, 깊이 영상 인코더에서 출력된 각각의 특징 지도가 융합 모듈의 입력으로 사용되며, 융합 모듈은 각 모달리티에서 중요한 정보를 추출한 후 보다 강건한 특징 지도를 생성한다. SGACNet[15]은 카메라 영상과 깊이 영상을 입력으로 받고 전역적 풀링(global pooling)을 진행한 후, 두개의 특징지도의 채널 축 어텐션 기법 융합 방법을 사용한다. 각 모달리티의 정보를 네트워크가 공유하기 때문에 높은 영상 분할 성능을 보이지만 단방향 융합 모델 대비 상대적으로 낮은 실시간성을 보인다.

Ⅲ. 이중 방향 융합 기법

3.1 영상 분할을 위한 다중 브랜치 네트워크

영상 분할은 픽셀 클래스 분류 결과를 출력 해야함과 동시에 객체의 형태, 주변 환경 등의 정보를 활용하여 분류 일관성을 유지하여야 한다. 이를 위하여 TBNs(Two Branch Networks) 구조는 디테일한 정보를 추출하는 공간 경로(Spatial Path)와 문맥적 정보를 추출하는 문맥 경로(Context Path)를 동시에 사용하는 방법을 제안하였다. TBNs 구조를 통해 픽셀 레벨의 정확도를 높임과 동시에 문맥적 정보를 학습할 수 있으며, 모든 구조가 단순한 합성곱 연산으로 구성되어 있어 실시간 동작이 가능하다. 그림 2는 영상 분할을 위한 TBN 모델의 구조를 나타낸다. 공간적 경로는 일반적인 합성곱 인코더의 구조를 사용하며, 입력 영상에서 공간적으로 중요한 특징을 추출한다. 문맥적 경로는 높은 풀링 비율을 통해 영상 전반의 문맥 정보를 함축한다. 그림 2에서는 공간적 경로에서 입력 영상의 크기를 4배만큼 축소하여 특징 지도를 추출하고 있으며, 문맥적 경로는 입력 영상 크기의 32배만큼 축소된 특징 지도를 추출하여 입력 영상의 함축적, 문맥적 특징 지도를 출력하는 모습을 표현하였다. 이후 공간적 경로와 문맥적 경로의 출력이 융합 및 디코더 모듈을 거치며 최종 분류 결과가 출력된다.

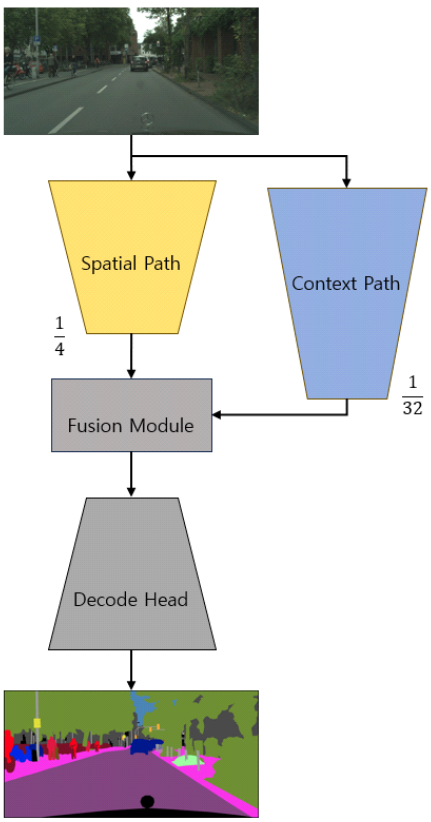


Fig. 2. Structure of TBNs

3.2 실시간 영상 분할을 위한 TBNs 구조에서의 이중 방향 융합 기법

본 논문에서 제안하는 기법의 전체적인 구조는 그림 3과 같다. 실시간 동작을 위해 복잡한 융합 기법 대신 단방향 융합 기법 구조를 사용하였으며, 카메라와 깊이 영상 각각의 공간적, 문맥적 특성을 활용하기 위하여 TBNs 구조가 사용되었다. 제안하는 기법에서의 특징 지도 융합 방법은 아래와 같다. 카메라 영상과 깊이 영상을 입력 받는 두 개의 TBN 구조 인코더를 통하여 각각의 특징 지도가 추출된 후 인코더 사이에 설계된 단방향 융합 모듈에 의해 특징지도가 융합된다. 깊이 영상은 넓은 영역에서 강건한 출력 특성을 보이므로 깊이 영상 인코더에서 카메라 인코더 방향으로 **문맥적 경로의 특징 지도**가 융합된다. 반대로 카메라 모델은 높은 밀도의 색 정보를 포함하므로 더 정밀한 공간적 정보를 포함한다. 따라서 **공간적 경로의 특징지도**는

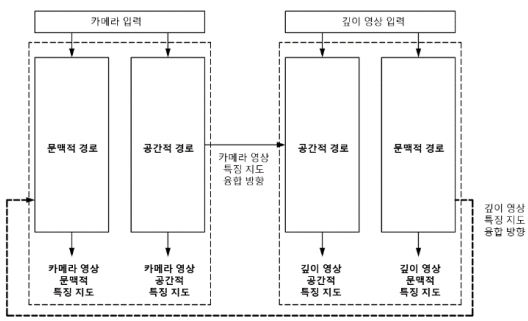


Fig. 3. Structure of the proposed method

카메라 인코더에서 깊이 영상 인코더로 융합된다. 융합 모듈은 단방향 방식으로 특징 지도를 융합하며, 공간적 경로와 문맥적 경로에서 입력 데이터에 따라 융합 방향이 반대로 설정되었다. 이러한 이중 방향 융합 기법은 단방향 융합 기법과 같은 연산량을 갖기 때문에 실시간 영상 분할에 적합한 융합 기법이며, 데이터 특성에 따른 융합 방향을 설정해 줌으로써 양방향 융합 기법의 장점을 모사하는 기능을 수행할 수 있다.

3.3 PIDNet 구조 기반 이중 방향 융합 구조

본 논문은 PIDNet 모델을 베이스라인 모델로 사용하였다. PIDNet은 영상 분할을 위해 TBNs 구조에 뎀퍼 브랜치 추가한 모델이다. 따라서 인코더는 3개의 특징 지도 추출 경로를 가지며, 각각 공간적 경로, 문맥적 경로 및 뎀퍼 경로로 명명된다. 뎀퍼 경로는 영상 분할 출력에서 넓은 영역이 작은 영역을 침범하는 문제를 해결하기 위해 제안되었다. 제안하는 논문에서 뎀퍼 경로는 융합을 진행하지 않았으며, 공간적 경로와 문맥적 경로에서 이중 방향 융합 모듈이 동작하도록 설계하였다.

Ⅳ. 실험 결과

4.1 실험 환경 및 학습 개요

실험은 모두 Cityscapes 데이터셋[10]을 사용하여 진행하였다. Cityscapes 데이터셋은 27개 도시에서 수집된 5000개 이미지로 구성 되어있다. 총 30개의 클래스가 정의되어 있으며, 19개의 클래스가 선정되어 영상 분할 대상으로 정의된다. 영상 분할 대상인 클래스는 도로, 인도, 건물, 벽, 펜스, 기둥, 교통 신호등, 교통 신호판, 식생, 수풀, 하 늘, 사람, 차, 트럭, 버스, 기차, 오토바이, 자전거, 이륜차 운전자로 구성된다.

모델 학습은 Nvidia Geforce RTX 3090을 사용하였으며, 미니 배치 사이즈는 12, 총 120,000 이터레이션으로 학습되었다.

제안하는 기법의 타당성을 비교하기 위해 6개의 모델을 학습하였다. 카메라(RGB), 깊이 영상(Depth)을 각각 입력으로 받는 2개의 모델을 학습하여 융합을 수행하지 않은 모델의 성능을 확인하였다. 본 논문에서 제안하는 기법(**BA: Bidirectional Addition**)과의 성능 비교를 위해 단방향 요소별 덧셈(EWA: Element-Wise Addition), 단방향 공간 어텐션(SA: Spatial Attention)을 학습하였다. 단방향 융합 기법은 기존의 연구들과 같이 깊이 특징 지도를 카메라 특징 지도 방향으로 융합하였으며, 가장 단순한 형태의 융합 기법으로 사용되는 융합 구조이다. 이중 방향 기법의 융합 방향 적절성을 평가하기 위하여 제안하는 기법의 방향을 정반대로 설계한 모델(BA opposite)을 학습하였다. 즉, BA opposite 모델에서 공간적 경로는 깊이 영상 인코더에서 카메라 인코더 방향으로, 문맥적 경로는 카메라 인코더에서 깊이 영상 인코더 방향으로 융합이 진행되었다.

4.2 성능 비교

학습한 결과들을 표 1에 정리하였다. 카메라, 깊이 영상을 입력으로 받는 모델의 경우 각각 79.96 mIoU, 58.29 mIoU의 성능을 보였다. 깊이 영상의 경우 스테레오 카메라로부터 얻어진 픽셀 상이(Disparity)를 사용하기 때문에 정확도가 낮으며, 색 정보가 존재하지 않기 때문에 카메라 모델 대비 21.67 mIoU 만큼 성능 하락을 보인다. EWA의 경우 카메라에 깊이 정보가 융합되어 카메라 모델 대비 0.82 mIoU 만큼의 성능 향상을 보이며, 멀티모달 융합에 의한 성능 향상을 확인할 수 있다. SA 모델은 어텐션 메커니즘을 통해 융합에 중요한 부분과 중요하지 않은 부분을 구분한 후 특징 지도를 업데이트 한다. 이를 통해 단순한 EWA 방식보다 0.43 mIoU의 추가적인 성능 향상을 얻을 수 있다. BA 기법의 경우 RGB 모델 대비 1.27 mIoU의 성능 향상을 보였으며, SA 기법보다 더 높은 성능을 보이는 결과를 확인할 수 있다. 제안하는 BA 기법은 어텐션 메커니즘을 사용하지 않으며, EWA 기법의 단순한 융합 기법을 사용하지만 SA 기법 대비 높은 성능을 보이는 결과를

Table 1. Performance Comparison

Modalities	road	sid.	bui.	wal.	fen.	pol.	lig.	sig.	veg.	ter.	sky	per.	rid.	car	tru.	bus	tra.	mot.	bic.	mIoU	FPS
RGB	98.2	85.9	93.2	56.9	66.6	67.4	73.9	81.6	92.7	64.8	94.8	83.8	68.0	95.6	80.7	90.0	78.5	66.8	79.4	79.96	-
Depth	96.8	74.2	83.7	30.3	43.4	48.6	41.3	47.1	80.4	42.0	83.8	65.0	46.4	88.5	49.1	58.9	49.8	27.6	50.5	58.29	-
RGB-D (Oppo)	98.3	85.8	93.1	54.9	66.1	65.6	74.1	80.3	92.7	64.7	95.4	83.8	68.5	95.4	83.8	91.5	84.9	67.2	79.4	80.31	14.49
RGB-D (EWA)	98.1	85.0	93.4	56.9	**70.7**	68.5	74.3	**81.9**	92.7	64.8	**95.2**	84.6	69.2	95.8	83.1	89.7	82.1	68.8	**79.9**	80.78	16.84
RGB-D (SA)	98.3	**86.4**	93.3	54.3	68.3	67.3	**74.7**	81.8	92.6	63.8	95.1	84.6	69.5	**96.0**	**88.9**	**92.3**	84.7	**71.4**	79.6	81.21	14.43
RGB-D (BA)	**98.4**	86.3	**93.5**	**61.1**	68.4	**67.5**	74.4	81.2	**92.8**	**64.9**	95.1	**84.7**	**69.6**	95.7	83.6	92.2	**85.7**	68.6	79.6	**81.23**	16.32

통해 제안하는 이중방향 융합 기법이 기존 기법 대비 높은 연산 효율성, 강건한 결과를 출력하는 것을 확인할 수 있다. BA opposite 기법은 80.31 mIoU의 성능을 보이며 비교 대상 융합 모델 중 가장 낮은 성능을 보였다. 가장 단순한 융합 모델인 EWA 모델보다도 낮은 성능을 보인 결과를 통해 공간적, 문맥적 특징지도 융합 방향의 적절한 설정이 매우 중요하다는 사실을 확인할 수 있다. 3개의 융합 모델의 FPS(Frames Per Seconds)가 표 1의 우측에 표현되어 있다. BA기법은 EWA와 비슷한 연산속도를 보였으며, SA 기법은 이보다 더 낮은 속도를 보였다. 이를 통해 본 논문이 제안하는 이중 방향 기법은 데이터의 특성을 활용하여 적절한 융합 방향을 설정함으로써 높은 연산 효율성과 높은 성능을 동시에 달성하는 것이 가능하다는 점을 제시한다.

V. 실험 분석

5.1 컨퓨전 매트릭스 정확도 분석

우선 학습된 모델의 픽셀별 정확도를 컨퓨전 매트릭스로 나타낸 결과를 분석하고자 한다. 깊이 영상 모델, 카메라 모델, BA 모델의 컨퓨전 매트릭스는 아래 그림 4~6에서 확인 가능하다. 세로 클래스가 정답 레이블이며, 가로 클래스는 모델이 출력한 분할 결과를 의미한다. 깊이 영상 모델의 경우(그림 4) 전반적으로 정확도가 낮은 모습을 보이며, 특히 비슷한 형태의 객체들 사이의 혼동된 결과를 확인 가능하다. 예를 들어 벽 클래스의 경우 39% 픽셀만이 정확히 분할되었고, 24%는 건물, 14%는 식생으로 잘못 분류하는 모습을 확인 가능하다. 이러한 양상은 자동차, 트럭 사이, 오토바이와 자전거 사이에서도 비슷한 양상을 보인다. 카메라 모델의 경우(그림 5) 색 정보를 사용하기 때문에 보다 정확한 분할 결과를 출력하는 모습을 확인 가능하다. 깊이 모델처럼 비슷한 형태의 클래스 오판이 줄어든 모습이 확인 가능하지만, 공간적 정부가 부족하여 벽을 건물로, 수풀을 식생으로 오판하는 경우가 많은 것을 확인 가능하다. 벽의 경우 19%를 건물로 잘못 판단하였으며, 수풀의 경우 11%를 식생으로 오판하는 모습을 확인할 수 있다. 제안된 모델의 경우(그림 6) 수풀과 식생의 오판 정도는 비슷하지만 벽을 건물로 오판하는 경우가 14%로 줄어든 결과를 관찰할 수 있다. 또한 오토바이를 포함한 여러 클래스에 대해서도 정확도가 향상된 모습을 확인할 수 있으며, 이를 통해 공간적 분별력이 향상된 결과를 확인 가능하다.

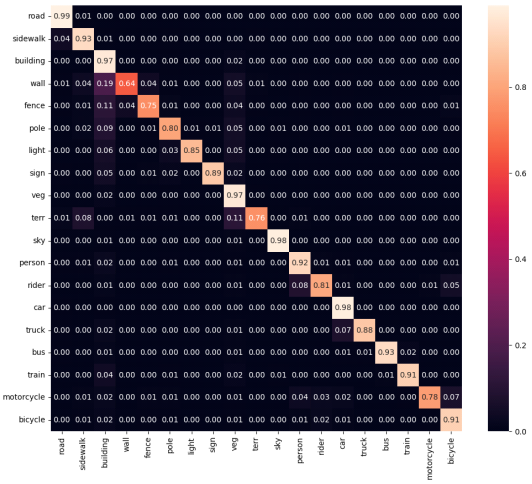


Fig. 5. Confusion Matrix of Camera Model Results

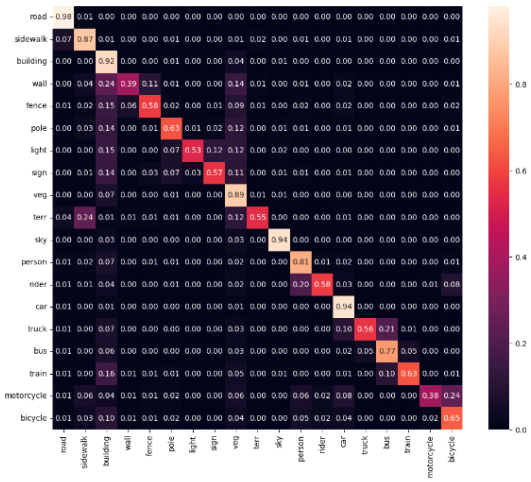


Fig. 4. Confusion Matrix of Depth Model Results

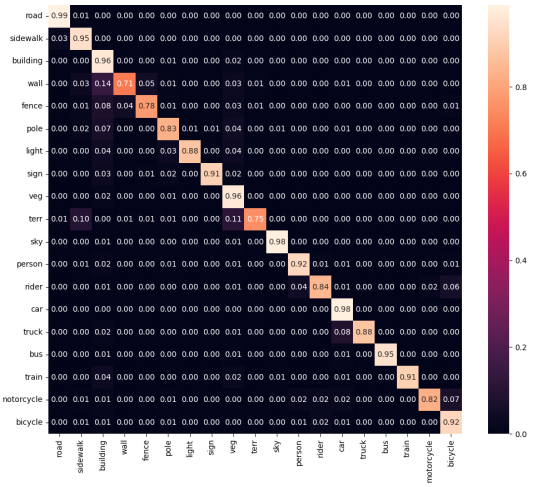


Fig. 6. Confusion Matrix of Proposed Method Results

5.2 출력 결과 정성적 평가

그림 7에는 정답 레이블, 카메라 모델의 출력, 제안하는 모델의 출력을 각각 시각화 하였다. 첫 번째 예시의 경우 정답 레이블의 붉은 박스 내 영역에서 자전거의 안장, 표지판 기둥 영역이 존재한다. 카메라 모델의 경우 자전거 안장 부분이 뒤편에 존재하는 자동차로 잘못 분할되었으며, 표지판 기둥 영역이 분할되지 않는 모습을 보인다. 깊이 정보를 활용한 모델의 경우 자전거 안장, 기둥이 모두 정확히 분할된 결과를 출력한다. 두 번째 예시의 경우 카메라 모델은 왼쪽 식생 영역을 정확히 구분해내지 못하는 반면 깊이 정보를 활용한 모델의 경우 해당 영역에 대한 분할 결과가 정확함을 확인 가능하다. 이러한 정량적 평가를 통해 제안하는 기법이 깊이 정보를 활용하여 보다 신뢰성 높은 영상 분할 결과를 보이는 것을 확인 가능하다.

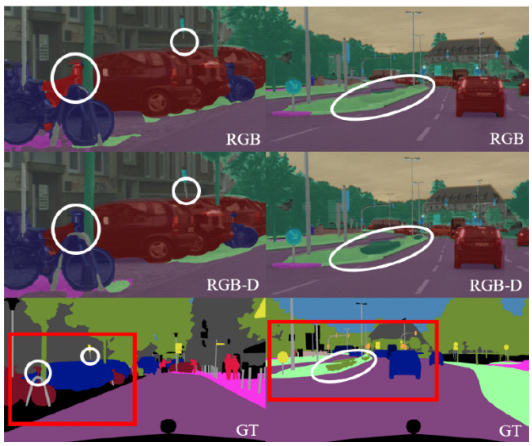


Fig. 7. Qualitative Results

Ⅵ. 결 론

본 논문은 영상 분할을 위한 새로운 다중센서 융합 기법을 제안한다. 제안하는 이중 방향 융합 기법은 영상 분할에서 카메라, 깊이 영상의 데이터 특성을 고려한 융합 방향 설계 철학을 기반으로 한다. 제안하는 기법을 통해 단방향 융합 방식보다 높은 영상 분할 성능을 얻을 수 있으며, 양방향 융합 방식보다 낮은 연산량이 요구되기 때문에 실시간 영상 분할을 위한 효율적인 모델 설계가 가능하다. 제안하는 이중 방향 융합 기법의 타당성을 검증하기 위해 정량적, 정성적 평가를 진행하였으며, 컨퓨전 매트릭스 비교를 통해 성능 향상 원인을 분석하였다.

References

[1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. CVPR*, pp. 2881-2890, 2017. (https://doi.org/10.1109/CVPR.2017.660)

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR*, pp. 770-778, 2016. (https://doi.org/10.1109/CVPR.2016.90)

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin "Attention is all you need," *Advances in NIPS*, 30, 2017. (https://doi.org/10.48550/arXiv.1706.03762)

[4] 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 NIPS*, 34, pp. 12077-12090, 2021. (https://doi.org/10.48550/arXiv.2105.15203)

[5] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. CVPR*, pp. 19529-19539, 2023. (https://doi.org/10.1109/CVPR52729.2023.01871)

[6] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 7088-7097, 2021. (https://doi.org/10.1109/ICCV48922.2021.00700)

[7] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proc. IEEE/CVF Conf. CVPR*, pp. 2869-2878, 2019. (https://doi.org/10.1109/CVPR.2019.00298)

[8] D. Seichter, S. B. Fischedick, M. Köhler, and H. M. Groß, "Efficient multi-task rgb-d scene analysis for indoor environments," in *2022 IJCNN*, pp. 1-10, Jul. 2022. (https://doi.org/10.1109/IJCNN55064.2022.9892852)

[9] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. CVPR*, pp. 12186-12195, 2022. (https://doi.org/10.1109/CVPR52688.2022.01187)

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. CVPR*, pp. 3213-3223, 2016. (https://doi.org/10.1109/CVPR.2016.350)

[11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541-551, 1989. (https://doi.org/10.1162/neco.1989.1.4.541)

[12] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, Y. Qiao, et al., "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. CVPR*, pp. 14408-14419, 2023. (https://doi.org/10.1109/CVPR52729.2023.01385)

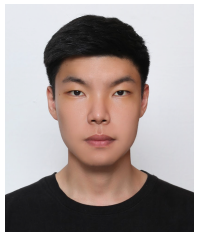
[13] Y. Shin, C. Lee, Y. Son, Y. Kim, J. Park, J. W. Choi, et al., "PIDDNet: RGB-Depth fusion network for real-time semantic segmentation," in *2023 14th Int. Conf. ICTC*, pp. 1049-1052, Oct. 2023. (https://doi.org/10.1109/ICTC58733.2023.10393276)

[14] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, pp. 325-341, 2018. (https://doi.org/10.48550/arXiv.1808.00897)

[15] Y. Zhang, C. Xiong, J. Liu, X. Ye, and G. Sun, "Spatial-information guided adaptive context-aware network for efficient RGB-D semantic segmentation," *IEEE Sensors J.*, 2023. (https://doi.org/10.1109/JSEN.2023.3304637)

[16] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robotics and Automat. Lett.*, vol. 5, no. 4, pp. 5558-5565, 2020. (https://doi.org/10.1109/LRA.2020.3007457)

신 윤 식 (Yunsik Shin)



2019년 2월 : 한양대학교 전기공학 학사
2019년 2월~현재 : 한양대학교 전기공학 석박통합과정
<관심분야> 딥러닝, 컴퓨터 비전
[ORCID:]0009-0000-2047-9205

손 용 호 (Yongho Son)

2022년 2월 : 스토니브룩 뉴욕주립대 컴퓨터공학 학사
2022년 2월~현재 : 한양대학교 인공지능학부 석박통합과정
<관심분야> 딥러닝, 컴퓨터 비전
[ORCID:]0009-0009-4114-1970

박 정 희 (Junghee Park)

1999년 2월 : 동서대학교 산업공학 학사
2006년 8월 : 성균관대학교 정보통신대학원 정보보호학 석사
2006년 7월~2008년 9월 : 코오롱베니트
2008년 10월~2009년 10월 : 한국정보인증
2009년 10월~2014년 2월 : DKUNC
2014년 2월~2016년 9월 : 케이엘넷
2016년 10월~현재 : LIG넥스원
<관심분야> 사이버전, 정보보안, PKI, 무인드론, 센서 융합
[ORCID:0000-0003-0334-6501]

이 채 현 (Chaehyun Lee)

2020년 2월 : 명지대학교 기계공학 학사
2022년 2월 : 명지대학교 기계공학 석사
2022년 2월~현재 : LIG넥스원 근무
<관심분야> 센서융합, 위치추정, 컴퓨터 비전, 자율주행
[ORCID:0000-0002-3363-1509]

김 양 곤 (Yanggon Kim)

2021년 2월 : 아주대학교 전자공학과 공학사
2023년 2월 : 아주대학교 AI융합네트워크학과 전자공학전공 공학사
2023년 1월~현재 : LIG 넥스원 지상통제연구소 연구원
<관심분야> 컴퓨터 비전, 영상처리, 딥러닝
[ORCID:0000-0003-2824-6654]

최 준 원 (Jun Won Choi)

2000년 2월 : 서울대학교 전기공학부 학사
2002년 2월 : 서울대학교 전기 및 컴퓨터 공학부 석사
2010년 5월 : 어바나-샴페인 일리노이주립대, 전기 및 컴퓨터 공학부 박사
2010년 8월~2013년 8월 : 퀄컴 (샌디에고) 근무
2013년 9월~2024년 2월 : 한양대학교 전기생체공학부 교수
2024년 3월~현재 : 서울대학교 전기정보공학부 교수
<관심분야> 로봇인지, 자율주행, 신호처리, 인공지능
[ORCID:0000-0002-3733-0148]