Journal of KIIT. Vol. 22, No. 5, pp. 13-22, May 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.5.13

교차 모달리티 캘리브레이션을 통한 RGB-깊이 영상 객체 분할

정애천*, 홍성은**

Cross Modal Calibration for RGB-D Instance Segmentation

Aecheon Jung*, Sungeun Hong**

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원 (IITP-2024-RS-2023-00254129) 연구결과와 정부(과학기술정보통신부)의 재원으로 한국연구재단-시스템반도체융합전문인력육성사업의 지원을 받아 수행된 연구임(2020M3H2A1078119)

요 약

딥러닝 분야에서 다중 모달리티 학습이 일반화되면서 이미지 분할 작업도 RGB뿐만 아니라 부가적인 모달리티를 함께 사용하는 추세이다. 특히, 센서 기술의 발전으로 과거에 비해 깊이 데이터를 쉽게 얻을 수 있게 되면서 깊이 데이터의 기하학적 정보를 활용하여 조명 변화에 민감한 RGB 기반 이미지 분할 모델의 취약점을 해결하려는 연구가 진행되고 있다. 그러나 기존의 RGB-깊이 연구는 주로 의미론적 분할 작업에 치중되어 있었다. 우리는 이러한 문제에 대응하기 위해 효과적인 RGB-깊이 객체 분할을 위한 융합 모듈을 제안한다. 또한, 세 가지 RGB-깊이 객체 분할 벤치마크 데이터셋을 새롭게 구축하고 다양한 기법들에 대해 비교실험을 수행한다. 이러한 데이터셋은 실내 탐색부터 로봇 조작까지 다양한 응용 분야를 지원할 수 있다. 제안하는 접근 방식은 다양한 객체 분할 작업 벤치마크 데이터셋에서 기존 기법 대비 뛰어난 인식 정확도를 보여준다.

Abstract

As multi-modal learning becomes more prevalent in the field of deep learning, image segmentation tasks are no longer limited to just RGB data but are increasingly incorporating additional modalities. Particularly, with advancements in sensor technology, it has become easier to obtain depth data, leading to research efforts aimed at leveraging the geometric information from depth data to address the vulnerability of RGB segmentation models to lighting variations. However, existing research and datasets in RGB-D primarily focus on semantic segmentation tasks. To overcome this limitation, we propose a powerful fusion module for RGB-D instance segmentation. Additionally, we have constructed three new benchmark datasets for RGB-D instance segmentation and conducted comparative experiments with various methods. These datasets support diverse applications, from indoor navigation to robotic manipulation. The proposed approach demonstrates superior accuracy compared to existing methods across various instance segmentation benchmark datasets.

Keywords

instance segmentation, rgb-d fusion, attention mechanism, segmentation benchmark dataset

* 인하대학교 전기컴퓨터공학과 석사과정
- ORCID: https://orcid.org/0009-0006-6736-0678
** 성균관대학교 실감미디어공학과 조교수(교신저자)
- ORCID: https://orcid.org/0000-0003-1774-9168

· Received: Mar. 20, 2024, Revised: Apr. 12, 2024, Accepted: Apr. 15, 2024
· Corresponding Author: Sungeun Hong
Dept. of Immersive Media Engineering, Sungkyunkwan University
Seoul, South Korea
Tel.: +82-2-740-1809, Email: csehong@skku.edu

Ⅰ. 서 론

최근 몇 년간 컴퓨터 비전 분야에서 딥러닝 기술의 발전과 더불어 시각 분할에 대한 관심이 크게 증가하고 있다. 시각 분할은 이미지를 픽셀 수준에서 의미적으로 관련된 영역으로 분할하는 작업을 의미하며 로봇 인식[1]과 같은 분야에 응용될 수 있다. 이 작업은 크게 의미론적 분할과 객체 분할로 분류할 수 있다. 이미지 내 픽셀들이 각각 어떤 클래스인지 분류하는 작업을 의미론적 분할이라고 하며 다른 객체라 해도 클래스가 같으면 이들을 구분하지 않는다. 이와 달리 객체 분할은 이미지 내 개별 객체들을 각각 구분하여 인지한다.

그림 1에서 확인 가능하듯이 객체 분할은 의미론적 분할과는 다르게 개별 의자들이 서로 구분되는 레이블을 가지는 것을 확인할 수 있다. 기존의 시각 분할 방법은 주로 RGB(Red, Green, Blue) 이미지를 사용하였지만 배경과 비슷한 색상과 질감을 가진 물체들을 구분하는 것은 어려운 문제이다. 특히, 조명 환경이 극단적인 경우에는 RGB 정보만으로 물체들을 식별하기 어렵다. 이에 따라 의미론적 분할 작업에서는 깊이 정보 및 기하학적 정보를 제공할 수 있는 깊이 데이터를 함께 활용하여 정확도를 높이려고 시도했다[2][3]. 깊이 정보 활용의 중요성에도 불구하고, 객체 분할 작업에서는 RGB-깊이 정보를 모두 활용한 연구가 미비했다. 이는 RGB-깊이 데이터셋에서 객체를 개별적으로 식별하고 분할하는 것이 중요한 문제임에도 불구하고, 적절한 벤치마크 데이터셋과 연구 방법이 부족했기 때문이다. 본 논문은 모달리티 사이의 공통적인 특성과 상호작용을 고려하여 객체 분할을 위한 교차 어텐션 기반의 융합 모듈을 제안한다. 또한, 실내 환경의 복잡성을 다루기 위해 기존 의미론적 분할에서 사용되던 RGB-깊이 데이터셋을 객체 분할 작업에 적합하게 변환시키고 나아가, 로봇 작업에 활용될 수 있는 상자 데이터셋을 새롭게 구축하였다.

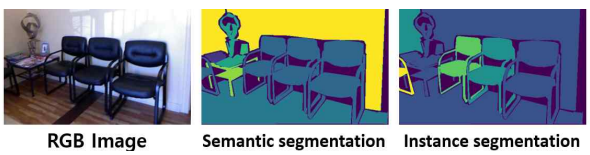


그림 1. 시각 분할 레이블 예시
Fig. 1. Examples of segmentation label

요약하면, 논문의 기여는 다음과 같다:

1. RGB-깊이 융합 모듈: RGB와 깊이 데이터가 공유하고 있는 정보를 기반으로 교차 어텐션을 통해 모달리티를 융합하는 효율적인 모듈을 제안한다.
2. RGB-깊이 객체 분할을 위한 데이터셋 구축: 실제 실내 환경을 반영한 객체 데이터셋 NYUDv2-IS와 SUN-RGBD-IS-kv2 및 생산 공정 환경에서 취득된 Box-IS 데이터셋을 새롭게 구축한다.
3. 포괄적인 성능 평가: 새로운 데이터셋과 여러 기법들에 대해 다양한 실험을 수행하여 후속 연구들을 위한 기준을 제시한다.

Ⅱ. 관련 연구

2.1 RGB-깊이 기반 객체 분할 알고리즘

객체 분할은 의미론적 분할과 객체 검출을 결합한 작업이다. 주요 목표는 이미지 내 개별 객체에 대한 픽셀 수준의 분할 마스크를 식별하고 그에 적절한 레이블을 할당하는 것이다. 이를 위해, 많은 딥러닝 모델들은 객체 감지 모델을 확장하여 객체를 먼저 감지한 후에 분할 작업을 수행한다. 특히, Mask-RCNN[4]은 Faster R-CNN[5]에 분할 작업을 위한 네트워크를 추가하여 객체 분할을 수행한다. 또한, Cascade Mask R-CNN[6]은 이 접근 방식을 발전시켜 다단계 모델 구조를 통합하여 성능을 향상시켰다. 이와 다르게 YOLACT[7] 및 CondInst[8]와 같은 모델들은 객체 감지와 분할을 병렬적으로 수행하여 더 높은 추론 속도를 달성하고 두 작업간의 관계를 밀접하게 고려했다. 최근 들어서는 ViT[9]와 같은 모델이 등장하면서 컴퓨터 비전 분야 내에서 트랜스포머 기반 접근 방식이 많이 시도 되고 있다. 특히, DETR[10], SOLQ[11]는 직접적인 집합 예측을 위해 이분 매칭 알고리즘을 사용하여 NMS와 같은 후처리 과정을 제거했다. 이러한 발전을 토대로 SOIT[12]은 ROI 자르기 과정을 제거하고 개별 객체에 대한 픽셀별 마스크를 직접 생성한다.

하지만, 객체 분할 작업에서 RGB 및 깊이 데이터를 함께 사용하는 경우는 일부 연구[13]를 제외하면 거의 진행되지 않았다. 이는 RGB-깊이 의미론적

분할 작업과는 다르게 RGB-깊이 객체 분할 작업에 특화된 데이터셋이 부족하기 때문이다. 이를 해결하기 위해 본 논문에서는 기존 RGB-깊이 의미론적 분할 작업에 사용되던 벤치마크 데이터셋인 NYUDv2[14]와 SUN-RGBD[15]를 COCO[16] 형식의 객체 분할 데이터셋으로 변환시켰다.

2.2 RGB-깊이 정보 융합 방법

RGB-깊이 융합 방법론은 주로 의미론적 분할 작업에서 사용되고 있는 방식이다. 초창기에는 RGB 및 깊이 정보를 결합하기 위해 단순히 원소별로 덧셈을 수행하는 방식을 사용했다[17]. 최근에는 다중 모달리티를 특징 수준에서 융합하는 기술을 많이 사용하고 있다. 이후 두 모달리티를 보다 효과적으로 융합하기 위해 채널 수준에서 모달리터 별 중요도에 따라 정보를 결합하는 방식[18], 비지역 접근법을 사용하는 방식[19], 융합 모듈을 통해 모달리티 사이의 차이를 다루고 문맥 정보 추출을 강화하는 방식[20], 그리고 RGB와 다른 모달리티 간의 공간 및 채널 차원에서의 상호작용을 고려하여 잡음을 줄이고 보완적 정보를 최대한 활용하는 방식[21] 등이 등장했다. 본 논문에서는 RGB 객체 분할 모델의 성능을 향상시키기 위해 두 모달리티가 공유하는 특징을 고려하면서 두 모달리티 사이의 상호작용을 통해 RGB와 깊이 데이터를 원활하게 통합하는 새로운 융합 모듈을 소개한다. 본 논문에서 제안하는 모듈은 단순히 두 모달리티 특징들을 원소별로 더해주는 FuseNet[17]과는 다르게 적응적으로 비중을 조절해서 융합한다. 또한, RGB 특징과 깊이 특징으로부터 각각에 대한 채널 별 중요도를 따로 구하는 ACNet[18] 방식과는 달리 융합된 특징으로부터 한 번에 중요도를 구함으로써 불필요한 작업을 줄였다. NANet[19] 방식은 컨볼루션 레이어를 통해 RGB, 깊이 특징들을 각각 공간적인 수준에서 수직 방향과 수평 방향으로 나누었다. 이를 서로 다른 모달리티끼리 교차하여 더해주는 방식으로 두 모달리티 사이의 상호작용을 다루었다. 이와 달리, 제안 모듈은 두 모달리티 사이의 교차 어텐션을 통해 상호작용을 직접적으로 모델링했다. CMX[21]의 경우에도 어텐션 기법을 사용하기는 했지만 어텐션 맵 구성시에 같은 모달리티끼리의 연산을 수행했다는 점에서 제안하는 모듈과 차이가 있다.

III. RGB-깊이 객체 분할을 위한 융합 모듈

본 논문에서는 서로 다른 모달리티 사이의 상호작용을 대칭적인 교차 어텐션을 통해 모델링하고 이를 채널 수준에서 중요도 기반으로 적응적으로 조절하는 간단하지만 효과적인 CMC(Cross Modal Calibration) 융합 모듈을 제안한다. RGB 정보와 깊이 정보는 각각 독립적인 네트워크를 통해 특징을 추출하고 그 사이에 CMC 융합모듈이 위치하여 두 모달리티 사이의 정보를 융합하는 역할을 하는 것을 그림 2를 통해 확인할 수 있다.

그림 3를 보면 융합모듈의 입력으로 들어가는 RGB와 깊이 특징 맵은 $F_{rgb}, F_d \in R^{C\times HW}$으로 표현되며 여기서 H와 W는 공간적인 차원을, C는 채널의 수를 의미한다.

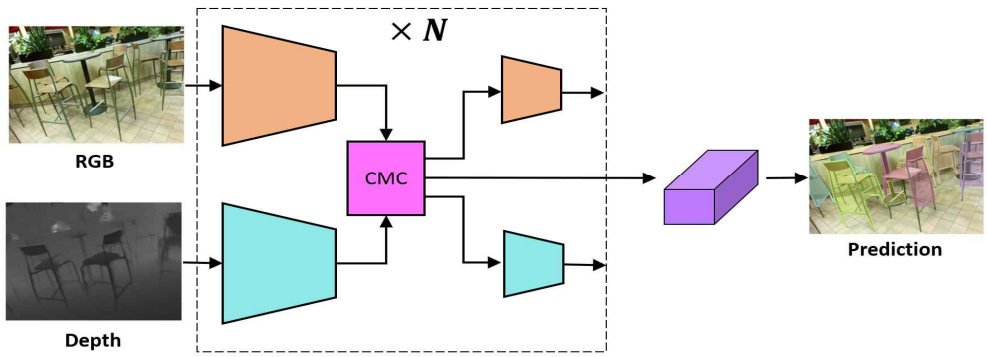


그림 2. 전체적인 모델 구조
Fig. 2. Overall model architecture

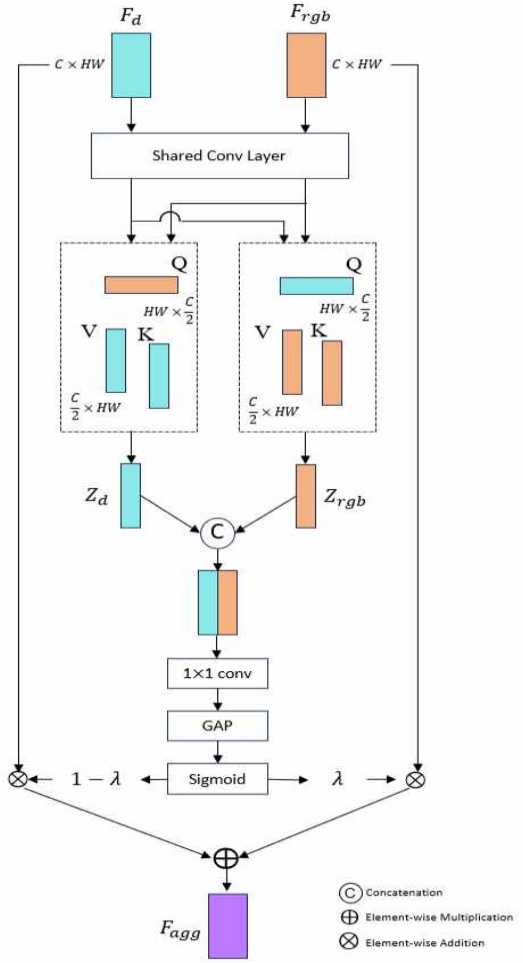


그림 3. RGB-깊이 융합 모듈
Fig. 3. RGB-D fusion module

각 모달리티 특징에 1×1 컨볼루션 레이어를 적용하여 쿼리(Query), 키(Key), 벨류(Value)를 구한다. 이때, RGB와 깊이 정보는 픽셀마다 서로 정렬되어 있는 동형 모달리티 관계다[22]. 이에 따라 쿼리, 키, 벨류 생성 시에 1×1 컨볼루션 레이어의 파라미터를 공유하여 두 모달리티 사이의 공통적인 특징을 고려하도록한다.

$$
\begin{aligned}
Q_{rgb\,or\,d} &= Conv_1(F_{rgb\,or\,d}) \in R^{\frac{C}{2}\times HW} \\
K_{rgb\,or\,d} &= Conv_2(F_{rgb\,or\,d}) \in R^{\frac{C}{2}\times HW} \\
V_{rgb\,or\,d} &= Conv_3(F_{rgb\,or\,d}) \in R^{\frac{C}{2}\times HW}
\end{aligned} \tag{1}
$$

식 (1)은 $C \times HW$ 크기의 RGB 특징 맵과 깊이 특징 맵을 1×1 컨볼루션 레이어에 통과시켜 $\frac{C}{2} \times HW$ 크기의 쿼리, 키, 벨류를 얻어낸다.

이렇게 얻어진 쿼리, 키, 벨류에 모달리티 사이의 상호작용을 반영하고 서로 구분되는 특성을 강조하기 위해 각 모달리티에서 얻어진 쿼리와 키를 서로 다른 모달리티 끼리 교차해서 곱한다. 이어서 소프트맥스를 적용하여 $HW \times HW$ 크기의 어텐션 맵을 구한다. 여기에 벨류를 곱한 후 컨볼루션 레이어를 통과시켜 식 (2)와 같이 RGB에 대한 특징을 담고 있는 $Z_{rgb} \in R^{\frac{C}{2}\times HW}$와 깊이에 대한 특징을 담고 있는 $Z_d \in R^{\frac{C}{2}\times HW}$를 얻을 수 있다. 이 두 텐서를 채널 방향으로 이어 붙여서 융합된 특징 $Z_{rgbd} \in R^{C\times HW}$가 나오게 된다.

$$
\begin{aligned}
Z_{rgb} &= Conv(softmax(Q_d^T K_{rgb})\, V_{rgb}^T) \\
Z_d &= Conv(softmax(Q_{rgb}^T K_d)\, V_d^T) \\
Z_{rgbd} &= Concat(Z_{rgb}, Z_d)
\end{aligned} \tag{2}
$$

융합된 특징을 1×1 컨볼루션과 전역적 평균 풀링을 통해 $R^{C\times 1}$ 크기의 벡터로 만들어 준 후 sigmoid 함수를 통해 정규화를 진행한다. 이렇게 얻어진 $R^{C\times 1}$ 크기의 벡터 λ는 두 모달리티 사이의 채널별 중요도를 나타내는 가중치 역할을 하게 된다. 이 가중치는 RGB 특징과 깊이 특징을 융합할 때 사용되며 기존의 RGB 특징과 깊이 특징들을 강화시켜서 다음 레이어로 전파시키기 위해 사용된다.

$$
\begin{aligned}
F_{agg} &= \lambda \otimes F_{rgb} + (1-\lambda) \otimes F_d \\
F_{rgb} &= F_{rgb} + (1-\lambda) \otimes F_d \\
F_d &= F_d + \lambda \otimes F_{rgb}
\end{aligned} \tag{3}
$$

식 (3)을 보면 RGB 특징과 깊이 특징의 채널에 대해 원소 별로 가중치 벡터 λ와 $1-\lambda$가 각각 곱해지게 되고 이를 더해서 최종적인 융합된 특징이 나오게 된다. 또한, RGB 특징에는 $1-\lambda$가 곱해진 깊이 특징을 더해주고 깊이 특징에는 λ가 곱해진 RGB 특징을 더해주는 방식으로 서로 다른 모달리티의 정보를 조금씩 반영하여 각각의 모달리티 특징들을 업데이트 해준다.

위 방식들을 통해 CMC 모듈은 기존 방식들과는 다르게 서로 다른 모달리티끼리 교차 어텐션을 직접 수행하여 두 모달리티 사이의 관계를 학습을 통해 효과적으로 계산한다.

또한, 채널 수준에서의 가중치를 각 모달리티 별로 구하지 않고 융합된 특징을 사용함으로써 중복되는 작업을 줄였다.

Ⅳ. 실험 결과 및 성능 평가

4.1 실험 데이터셋

합성 데이터셋이 아닌 실제 실내 환경에 대한 RGB-깊이 객체 분할 작업 데이터셋이 부족했기 때문에 아래와 같은 데이터셋들을 직접 구축하여 CMC 모듈의 유효성을 검증하기 위해 사용했다.

4.1.1 NYUDv2-IS

NYUDv2[14]은 RGB-깊이 의미론적 분할 작업에서 널리 쓰이고 있는 실내 장면들에 대한 벤치마크 데이터셋이다. 깊이 데이터는 Kinect v1센서로 취득되어졌으며 총 1,449장으로 이루어져 있다. 이를 객체 분할 작업을 위한 NYUDv2-IS 데이터셋으로 변환하기 위해 기존 데이터셋의 메타데이터를 활용하여 각 이미지 내의 물체들끼리 구분되는 마스크들을 추출하고 이를 COCO 형식으로 어노테이션으로 정제했다. 이 과정을 통해 각 물체들에 대해 이진 마스크를 다각형으로 변환하고 그에 대응되는 카테고리, 면적, 경계 상자, 분할 마스크를 부여하게 된다. 다만, 기존 NYUDv2 데이터셋은 물체가 아닌 카테고리도 포함되어 있었기 때문에 이는 제외한다. 이에 따라, NYUDv2-IS는 원래 있던 13개 중 9개의 카테고리만을 가지며 1,433장으로 구성되어 있다. 실험 시에는 의미론적 분할 작업에서 사용했던 설정과 동일하게 788장을 학습 데이터셋으로, 645장을 평가 데이터셋으로 사용했다.

4.1.2 SUN-RGBD-IS-kv2

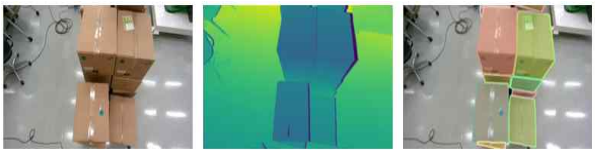
SUN-RGBD[15]는 다양한 실내 장면들을 담고 있는 대규모의 RGB-깊이 벤치마크 데이터셋이다. Intel Realsense, Asus Xtion, Kinect v1, Kinect v2와 같은 센서들로 깊이 데이터를 취득했으며 총 10,335장으로 구성되어 있다. 이를 객체 분할 작업에 맞는 형식으로 변환하기 위해 NYUDv2-IS에 사용된 방법을 유사하게 사용했으며 Kinect v2를 통해 취득한 샘플들만 모아 총 3,784장의 SUN-RGBD-IS-kv2 데이터셋을 구축했다. 그 과정에서 37개의 카테고리 중 17개만을 사용했다. 실험 시에는 3,784장을 랜덤하게 나누어 2,838장을 학습 데이터셋으로, 946장을 평가 데이터셋으로 사용했다.

4.1.3 Box-IS

Box-IS 데이터셋은 로봇이 상자를 옮기는 작업 등에 사용될 수 있는 RGB-깊이 객체 분할 데이터셋이다. Intel RealSense D455로 데이터를 취득했으며 총 543장으로 구성되어 있고 카테고리는 상자 1개이다. 스테레오 이미지에 UniMatch[23] 기법을 적용하여 고품질의 깊이 데이터를 얻었으며 상자가 잘 정렬되어 있는 쉬운 경우부터 상자가 임의로 쌓여있는 매우 복잡한 경우까지 다양한 상황을 다루고 있다. Box-IS 또한 COCO 형식의 어노테이션으로 이루어져 있고 경계 상자, 분할 마스크 등의 정보를 포함하고 있다. 실험 시에는 488장을 학습 데이터셋으로, 55장을 평가 데이터셋으로 사용했다.

(a) NYUDv2 객체 분할 데이터셋
(a) NYUDv2-IS

(b) SUN-RGBD-kv2 객체 분할 데이터셋
(b) SUN-RGBD-IS-kv2



(c) 상자 객체 분할 데이터셋
(c) Box-IS

그림 4. 데이터셋 예시
(왼쪽: RGB / 가운데: 깊이 / 오른쪽: 정답)
Fig. 4. Dataset examples
(Left: RGB / Middle: depth / Right: ground truth)

4.2 비교 기법

본 논문은 제안 기법의 성능을 비교하기 위해 RGB-깊이 데이터에 대한 여러 융합 방식들에 대한 실험을 수행했다. 초기 융합 기법은 별도의 융합 모듈 없이 RGB와 깊이 데이터를 채널 방향으로 이어 붙인 후 RGB 기반 객체 분할 모델에 입력으로 넣어주는 방식이다. 후기 융합 기법은 RGB와 깊이 데이터를 별도의 네트워크에서 처리하고 맨 마지막 레이어에서 융합하는 방식이다. 내부 어텐션 기반의 모듈은 각 모달리티 내부의 상호작용에 집중하는 방식이며 외부 어텐션 기반의 모듈은 각 모달리티들을 따로 처리하여 모달리티 사이의 상호작용을 이용하는 방식이다. 또한, 기존 의미론적 분할 작업에서 상위권의 성능을 보여주는 SA-Gate[24]와 CMX[21] 모듈을 객체 분할 모델에 적용해 비교기법으로 사용한다.

4.3 구현 상세 내역

기존의 RGB 객체 분할 모델의 백본 네트워크를 병렬적으로 배치하여 하나는 RGB 정보를 처리하도록 하고 나머지 하나는 깊이 정보를 처리하도록 했다. CMC 융합 모듈은 두 백본 네트워크 레이어 사이에 추가되어 RGB와 깊이 모달리티 사이의 정보를 교환하는 역할을 했다. 이때, 물체 탐지와 분할을 나눠서 수행하는 DETR[10] 모델과 한 번에 수행하는 SOLQ[11]모델을 실험에 사용했다.

4.4 정량적 평가

우리가 구축한 데이터셋들에 대해 다양한 융합 기법들을 적용해 비교 실험을 진행했다. 모든 실험에 대해 RGB 데이터만 사용한 경우보다 제안하는 융합 모듈을 사용했을 때 일관되게 높은 성능을 보여준다. 표 1은 NYUDv2-IS 데이터셋에 대한 실험 결과를 보여주고 있다. DETR과 SOLQ 모델에 CMC 모듈을 적용했을 경우 RGB 모델 대비 각각 2.0%, 3.0%의 AP 향상을 보여주고 있다. 또한, CMC 모듈이 다른 융합 기법들보다 높은 성능을 보여주고 있다.

표 1. NYUDv2-IS 데이터셋에 대한 성능 비교
Table 1. Performance comparison on NYUDv2-IS

	Method	AP	$AP_{0.5}$	$AP_{0.7}$	AP_S	AP_M	AP_L
DETR	RGB	31.2	53.8	31.7	5.3	18.9	41.1
	Initial fusion	25.6	45.4	25.9	4.1	14.6	34.3
	Final fusion	31.9	53.9	32.5	5.2	19.5	41.7
	Internal attention	29.5	51.6	29.4	5.7	18.2	38.9
	External attention	30.1	52.3	30.4	3.8	19.0	39.4
	SA-Gate[24]	30.7	52.5	31.2	4.2	19.2	40.2
	CMX[21]	31.4	54.2	31.9	4.6	20.7	40.4
	Ours	33.2	55.2	34.6	5.0	21.4	43.0
SOLQ	RGB	33.1	52.8	34.7	3.2	20.6	44.5
	Initial fusion	28.6	47.8	29.9	2.5	16.3	38.9
	Final fusion	34.9	55.5	37.5	5.6	21.5	46.5
	Internal attention	35.3	56.0	37.6	3.6	23.9	46.9
	External attention	35.1	55.5	37.4	4.5	23.5	45.6
	SA-Gate[24]	35.4	56.1	38.1	3.8	25.1	46.2
	CMX[21]	32.5	53.3	34.6	4.6	20.4	43.4
	Ours	36.1	57.3	38.5	4.8	23.9	46.6

초기융합을 했을 경우에는 RGB와 깊이 데이터를 전부 사용함에도 불구하고 RGB만 사용하는 경우보다 성능이 크게 떨어지며 DETR 모델에서는 다른 기법들도 성능이 낮게 나오는 것을 볼 수 있다. 이는 RGB와 깊이를 함께 쓰는 것이 무조건적인 성능 향상을 보장해주지는 않음을 시사한다. 또한, 의미론적 분할 작업에서 높은 성능 향상을 이끌어냈던 SA-Gate와 CMX 모듈들이 종종 RGB 모델보다도 못한 성능을 보여주고 있다. 이는 의미론적 분할과 객체 분할이 비슷한 작업이기는 해도 RGB와 깊이를 융합할 때 각 작업에 알맞은 방식을 활용해야 된다는 것을 암시한다.

표 2는 SUN-RGBD-IS-kv2 데이터셋에 대한 실험 결과를 보여주고 있다. NYUDv2-IS 데이터셋과 전체적인 경향성은 비슷하나 우리의 모듈을 적용했을 경우 NYUDv2-IS 데이터셋 비해 성능 향상의 정도가 적음을 확인할 수 있다. 이는 SUN-RGBD-IS-kv2 데이터셋이 클래스의 수가 더 많은 복잡한 데이터셋이기 때문으로 보인다. 표 3은 Box-IS 데이터셋에 대한 결과이며 대부분의 기법들이 매우 높은 성능을 보여주고 있다. 특히, 제안하는 모듈을 적용한 경우, 두 번째로 높은 성능을 보여준 내부 어텐션 기법보다 약 0.4% 높은 성능을 보인다.

표 2. SUN-RGBD-IS-kv2 데이터셋에 대한 성능 비교
Table 2. Performance comparison on SUN-RGBD-IS-kv2

	Method	AP	$AP_{0.5}$	$AP_{0.7}$	AP_S	AP_M	AP_L
SOLQ	RGB	20.4	31.5	20.8	1.1	9.7	28.6
	Initial fusion	19.4	30.2	19.9	1.0	8.4	27.3
	Final fusion	21.5	33.0	22.4	1.3	10.5	30.0
	Internal attention	22.6	35.5	23.2	2.1	11.8	31.0
	External attention	22.4	34.2	23.5	2.1	11.3	30.9
	SA-Gate[24]	21.6	33.6	22.9	1.2	12.6	29.8
	CMX[21]	18.6	30.0	19.5	2.8	8.1	26.6
	Ours	22.9	35.7	24.1	3.3	15.6	31.0

표 3. Box-IS 데이터셋에 대한 성능 비교
Table 3. Performance comparison on Box-IS

	Method	AP	$AP_{0.5}$	$AP_{0.7}$	AP_S	AP_M	AP_L
SOLQ	RGB	83.3	93.5	87.6	7.9	54.0	87.9
	Initial fusion	81.6	92.6	86.4	1.8	52.0	86.9
	Final fusion	83.2	92.9	87.0	31.5	55.1	87.8
	Internal attention	83.4	94.6	87.7	14.0	56.3	87.8
	External attention	83.3	93.2	87.1	14.6	55.2	88.0
	Ours	83.8	93.8	87.8	14.4	52.6	88.4

4.5 정성적 평가

그림 5는 여러 데이터셋 샘플들에 대해 다양한 기법들로 추론을 해보고 그 결과를 특정 객체에 대해 시각화한 결과이다. 1~3행은 NYUDv2-IS, 4~5행은 SUN-RGBD-IS-kv2, 6행은 Box-IS 데이터셋에 대한 결과이며 빨간 상자는 우리의 모듈이 다른 기법들에 비해 더 좋은 분할 성능을 보여주는 부분을 강조한 것이다. 전체적으로 우리의 모듈을 사용했을 경우에 물체의 말단 부분을 잘 인지하고 분할 마스크를 예측한 것을 확인할 수 있다. 특히, 6행에서는 물체가 가려진 상황에서도 다른 기법들에 비해 정확한 마스크를 추론하는 것을 확인할 수 있다.

V. 결 론

본 논문에서는 기존 RGB 모델이 가지고 있던 한계를 극복하기 위해 깊이 정보를 활용하는 것의 중요성에 대해 다루었다. 특히, 효과적인 RGB-깊이 객체 분할 작업을 위해 CMC(Cross Modal Calibration) 융합 모듈을 제안했다.

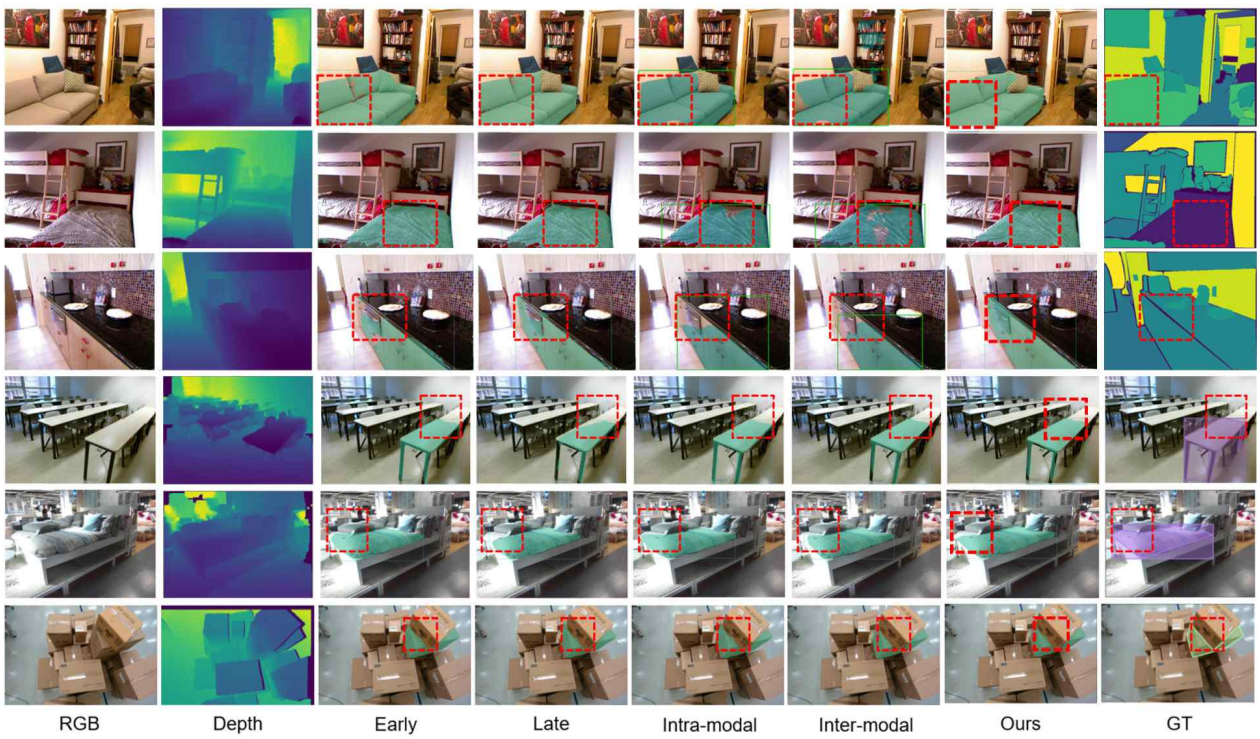


그림 5. NYDv2-IS, SUN-RGBD-IS-kv2, Box-IS 데이터셋에 대해 적용된 여러 기법들의 시각적인 비교
Fig. 5. Visual comparison of various methods applied to the NYDv2-IS, SUN-RGBD-IS-kv2, and Box-IS

이 모듈은 파라미터 공유 전략을 통해 모달리티 사이의 공통 특성들을 다루고 교차 어텐션을 통해 모달리티끼리의 연관 관계를 파악하여 두 모달리티를 채널 수준에서의 중요도를 기반으로 융합하는 간단하지만 효과적인 방식을 취한다. 또한 이러한 RGB-깊이 객체 분할 작업에 대해 잘 구성된 3가지 벤치마크 데이터셋을 소개한다. NYUDv2-IS와 SUN-RGBD-IS-kv2는 기존의 실내 환경에 대한 의미론적 분할 데이터셋을 재가공하여 얻어졌으며 Box-IS는 실제 공장 환경에서 상자가 쌓여있는 다양한 시나리오들에 대해 직접 취득되었다. 이러한 데이터셋은 기존에 RGB-깊이 객체 분할연구의 한계를 극복할 수 있으며, 로봇 조작, 인간 보조 작업 등 다양한 응용분야에서의 발전을 위한 청사진을 제시한다. 이에 더해 다양한 기법과 데이터셋에 대한 실험을 통해 본 논문은 RGB-깊이 객체 분할 분야에 진입하려는 연구자들에게 베이스라인 제공 및 연구의 초석을 제공한다는 점에서 의의가 있다.

References

[1] J. Li, Y. Dai, J. Wang, X. Su, and R. Ma, "Towards broad learning networks on unmanned mobile robot for semantic segmentation", 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, pp. 9228-9234, May 2022. https://doi.org/10.1109/ICRA46639.2022.9812204.

[2] 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", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 7088-7097, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00700.

[3] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation", Proc. of European Conference on Computer Vision (ECCV), Munich, Germany, Vol. 11215, pp. 144-161, Oct. 2018. https://doi.org/10.1007/978-3-030-01252-6_9.

[4] K. He, G. Gkioxari, . Dollár, and R. Girshick, "Mask r-cnn", 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2961-2969, Oct. 2017. https://doi.org/10.1109/ICCV.2017.322.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", arXiv:1506.01497 [cs.CV], Jun. 2015. https://doi.org/10.48550/arXiv.1506.01497.

[6] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 5, pp. 1483-1498, May 2021. https://doi.org/10.1109/TPAMI.2019.2956516.

[7] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, pp. 9157-9166, Oct. 2019. https://doi.org/10.1109/ICCV.2019.00925.

[8] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation", Proc. of European Conf. on Computer Vision (ECCV), Glasgow, UK, Vol. 12346, pp. 282-298, Aug. 2020. https://doi.org/10.1007/978-3-030-58452-8_17.

[9] A. Dosovitskiy, et al, "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv:2010.11929 [cs.CV], Oct. 2020. https://doi.org/10.48550/arXiv.2010.11929.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers", Proc. of European Conference on Computer Vision (ECCV), Glasgow, UK, Vol. 12346, pp. 213-229, Aug. 2020. https://doi.org/10.1007/978-3-030-58452-8_13.

[11] B. Dong, F. Zeng, T. Wang, X. Zhang, and YichenWei, "Solq: Segmenting objects by learning queries", Proc. of Neural Information Processing Systems (NeurIPS), Vol. 34, pp. 21898-21909, 2021.

[12] X. Yu, D. Shi, X. Wei, Y. Ren, T. Ye, and W. Tan, "Soit: Segmenting objects with instance-aware transformers", Proc. of Int'l Conference on Artificial Intelligence (AAAI), Pennsylvania State University, USA, Vol. 36, pp. 3188-3196, Feb. 2022. https://doi.org/10.1609/aaai.v36i3.20227.

[13] M. Wang, L. Hu, Y. Bai, X. Yao, J. Hu, and S. Zhang, "Amnet: a new rgb-d instance segmentation network based on attention and multi-modality", The Visual Computer, Vol. 40, pp. 1-15, 2024. https://doi.org/10.1007/s00371-023-02850-w.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images", Proc. of European Conference on Computer Vision (ECCV), Florence, Italy, Vol. 7576, pp. 746-597, Oct. 2012. https://doi.org/10.1007/978-3-642-33715-4_54.

[15] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite", Proc. of Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 567-576, Jun. 2015. https://doi.org/10.1109/CVPR.2015.7298655.

[16] T.-Y. Lin, et al, "Microsoft coco: Common objects in context", Proc. of European Conference on Computer Vision (ECCV), Zurich, Switzerland, Vol. 8693, pp. 740-755, Sep. 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

[17] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture", Proc. of Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, Vol. 10111, pp. 213-228, Nov. 2016. https://doi.org/10.1007/978-3-319-54181-5_14.

[18] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation", IEEE Int'l Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 144-1444, Sep. 2019. https://doi.org/10.1109/ICIP.2019.8803025.

[19] G. Zhang, J.-H. Xue, P. Xie, S. Yang, and G. Wang, "Non-local aggregation for rgb-d semantic segmentation", IEEE Signal Processing Letters, Vol. 28, pp. 658-662, Mar. 2021. https://doi.org/10.1109/LSP.2021.3066071.

[20] E. Yang, W. Zhou, X. Qian, and L. Yu, "Mgcnet: Multilevel gated collaborative network for rgb-d semantic segmentation of indoor scene", IEEE Signal Processing Letters, Vol. 29, pp. 2567-2571, Dec. 2022. https://doi.org/10.1109/LSP.2022.3229594.

[21] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers", IEEE Transactions on Intelligent Transportation Systems, Vol. 24, No. 12, pp. 14679-14694, Dec. 2023. https://doi.org/10.1109/TITS.2023.3300537.

[22] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers", Proc. of Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 12186-12195, Jun. 2022. https://doi.org/10.1109/CVPR52688.2022.01187.

[23] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 11, pp. 13941-13958, Nov. 2023. https://doi.org/10.1109/TPAMI.2023.3298645.

[24] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation", Proc. of European Conference on Computer Vision (ECCV), Glasgow, UK, Vol. 12356, pp. 561-577, Aug. 2020. https://doi.org/10.1007/978-3-030-58621-8_33.

저자소개

정 애 천 (Aecheon Jung)

2022년 2월 : 인하대학교 기계공학과(공학사)
2022년 8월 ~ 현재 : 인하대학교 전기컴퓨터공학과 석사과정
관심분야 : Multimodal learning, Missing modality, Parameter efficient learning

홍 성 은 (Sungeun Hong)

2010년 2월 : 한양대학교 컴퓨터공학과(공학사)
2012년 8월 : 카이스트 전산학과(공학석사)
2018년 2월 : 카이스트 전산학과(공학박사)
2018년 1월 ~ 2020년 8월 : SK telecom (T-Brain, AI Center) 연구원
2020년 9월 ~ 2023년 8월 : 인하대학교 정보통신공학과 조교수
2023년 9월 ~ 현재 : 성균관대학교 실감미디어공학과 조교수
관심분야 : Domain Adaptation, Multimodal Learning, Face Understanding, Video Object Segmentation