

논문 2026-3-18 <http://dx.doi.org/10.29056/jsf.2026.09.18>

# 포인트 클라우드에서 복셀화 기법에 따른 3D CNN 학습 모델의 XAI 성능 비교 분석

김재홍\*, 서재석\*\*, 박기욱\*\*\*, 김광수\*\*\*\*, 이기훈\*\*\*\*\*†

## Comparative Analysis of XAI Performance in 3D CNN Models with Different Point Cloud Voxelization Methods

Jaehong Kim\*, Jaeseok Seo\*\*, Kiwook Park\*\*\*, Kwangsoo Kim\*\*\*\*, Kihoon Lee \*\*\*\*\*†

### 요 약

본 논문에서는 도심 포인트 클라우드 의미론적 분할을 수행하는 Dense 3D CNN을 대상으로 입력 채널 기여도와 공간적 활성화를 설명가능인공지능(XAI) 관점에서 분석하였다. STPLS3D 일부 영역을 크롭·복셀화하고, occupancy, normalized relative z-height, R16, G16, B16으로 구성된 Full-Z+RGB-16 입력을 사용하여 GradientSHAP, Grad-CAM, ROAR를 적용하였다. 실험 결과 Building과 Car 클래스 모두에서 occupancy의 평균 절대 SHAP 기여도가 가장 높았고, 높이 및 색상 채널의 상대적 중요도는 클래스에 따라 다르게 나타났다. Grad-CAM에서는 실제 점유 복셀뿐 아니라 패치 경계에서도 활성화가 관찰되어 경계 효과 가능성을 확인하였으며, ROAR에서는 40% 채널 제거 시 SHAP 순위 제거의 Car IoU가 20.79%로 무작위 제거의 36.24%보다 약 15.45%p 낮게 나타났다.

### Abstract

This paper analyzes input channel contributions and spatial activations of a dense 3D CNN for urban point cloud semantic segmentation from an Explainable AI (XAI) perspective. A cropped region of the STPLS3D dataset was voxelized using the Full-Z+RGB-16 representation, consisting of occupancy, normalized relative z-height, R16, G16, and B16, and GradientSHAP, Grad-CAM, and ROAR were applied. The results showed that occupancy had the highest mean absolute SHAP attribution for both Building and Car, while the relative importance of height and color channels differed by class. Grad-CAM activations were observed around occupied voxels as well as near patch boundaries, indicating a possible boundary effect. In the ROAR experiment, at a 40% channel removal rate, the Car IoU after SHAP ranked removal was 20.79%, approximately 15.45 percentage points lower than the 36.24% obtained with random removal.

**한글키워드 :** 3D CNN, 복셀화, 설명가능한 인공지능, 의미론적 분할

**keywords :** 3D CNN, Voxelization, Explainable AI (XAI), Semantic Segmentation

\* 동국대학교 컴퓨터AI학부

\*\* 건양대학교 의료인공지능학과

\*\*\* 울산대학교 IT융합학부

\*\*\*\* 한국전자통신연구원 퍼지컬AI연구소

† 교신저자:이기훈(email: hoony@etri.re.kr)

접수일자: 2026.09.12. 심사완료: 2026.09.19.

게재확정: 2026.09.20.

## 1. 서론

도심 포인트 클라우드의 의미론적 분할 성능은 포인트를 복셀로 변환하는 방법과 입력 특징의 구성에 따라 달라질 수 있다. 최근 수행된 연구에서는 STPLS3D 데이터셋을 대상으로 Baseline, RGB-16, Full-Z 및 Full-Z + RGB-16의 네 가지 복셀 표현을 Sparse 3D CNN에 적용하였다. 그 결과, 수평 패치 크기 32 조건에서 Full-Z+RGB-16이 가장 높은 mIoU를 기록하였다 [1]. 이 결과는 수직 방향의 연속성을 유지하는 공간 분할 방식과 복셀 내부의 지배적 색상을 보존하는 대표값 선정 방식이 분할 성능에 영향을 줄 수 있음을 보여준다.

그러나 선행연구는 복셀 표현별 예측 성능을 비교하는 데 중점을 두었으며, 모델이 형상, 높이 및 색상 중 어떤 정보를 주요 근거로 활용했는지에 대한 분석은 충분히 이루어지지 않았다. 분할 성능이 향상되더라도 그 원인이 특정 입력 채널의 정보량 증가에 있는지, 공간적 문맥의 보존에 있는지, 또는 데이터와 모델 구조의 상호작용에 있는지는 성능지표만으로 판단하기 어렵다. 따라서 모델의 예측 근거를 입력 채널과 공간 영역의 두 측면에서 분석할 필요가 있다.

본 논문은 선행논문의 색상 표현을 참고하여 점유·상대 높이·RGB-16의 5채널 표현을 분석 대상으로 설정하고, Dense 3D U-Net에 GradientSHAP(Gradient Shapley Additive exPlanations), Grad-CAM(Gradient-weighted Class Activation Mapping) 및 ROAR(RemOve And Retrain)의 개념을 적용한 채널 제거 후 재학습 실험을 수행한다. GradientSHAP은 점유, 높이 및 RGB 채널이 클래스별 예측에 미친 기여도를 정량화하며, Grad-CAM은 모델의 공간적 활성화 위치를 시각화한다. 채널 제거 후 재학습 실험은 SHAP이 중요하다고 평가한 채널을 제거

했을 때 성능이 실제로 감소하는지 확인하는 데 사용한다 [2], [3], [4].

이러한 배경에서 본 논문은 두가지 목적을 달성하기 위해 연구를 수행하고자 한다. 첫 번째로, 점유·상대 높이·RGB-16의 5채널 입력을 사용하는 Dense 3D CNN이 어떤 입력 정보와 공간 영역에 의존하는지를 탐색하고자 한다. 두 번째로, 제한된 채널 수와 패치 규모에서 설명가능성 결과를 해석할 때 발생할 수 있는 문제를 제시하는 것을 목적으로 한다.

## 2. 관련 연구

Sparse 3D CNN과 같이 점유된(occupied) 복셀에 대해서만 연산을 수행하는 구조는 일반적인 dense CNN과 달리 고정된 격자 형태의 feature map을 전제로 하지 않기 때문에, 표준 Grad-CAM에서 사용하는 convolutional feature map과 gradient를 동일한 방식으로 추출·해석하기 어렵다는 한계가 있다. 최근 point cloud semantic segmentation 분야에서는 이러한 한계를 고려하여 gradient 기반 localization 기법을 point 단위로 확장하려는 연구가 이루어지고 있다 [5]. 그러나 voxel 기반 sparse 구조에 대한 표준화된 설명가능성 적용 방법은 아직 충분히 정립되지 않은 상황이다.

3D 딥러닝 전반을 조망한 최근 연구들은 point cloud 표현학습과 분할 모델의 성능 향상에 관한 연구가 빠르게 축적되어 온 것과 대조적으로, 모델이 어떠한 입력 정보를 근거로 예측 결과를 도출하는지를 규명하는 해석 연구는 여전히 초기 단계에 머물러 있음을 지적하고 있다 [7, 8]. 최근에는 point cloud 분류 모델에 다중 인스턴스 학습 기반의 설명가능성을 적용한 연구도 발표되고 있다 [9]. 이처럼 point cloud와 XAI를 결합한 연구가 지속적으로 이루어지고 있으나, 기존 연

구의 상당수는 분류(classification)에 집중되어 있으며, 의미론적 분할(semantic segmentation), 특히 voxel 입력 채널별 기여도를 정량적으로 분석한 연구는 상대적으로 부족하다.

기존 연구에서는 도메인 지식을 뉴런의 중요도와 연계함으로써 모델의 의사결정 과정을 채널 단위에서 분석할 수 있음을 제시하였다 [10]. 본 논문에서 occupancy, z-height, RGB 등 voxel 입력 채널별 SHAP 기여도를 비교하는 접근 역시 입력 정보를 채널 단위로 구분하여 모델의 예측 근거를 분석한다는 점에서 이러한 채널 단위 중요도 개념과 개념적 연관성을 가진다.

한편 CAM 계열 기법을 분할 문제에 적용할 경우, 분류 문제와 달리 공간적 경계와 활성화 영역 간 관계를 해석하는 데 주의가 필요하다는 점이 최근 연구에서도 지적된 바 있다 [11]. 이는 본 논문에서 Grad-CAM activation이 실제 점유 영역뿐만 아니라 patch 경계 부근에서도 나타난 현상과 관련되며, zero-padding에 따른 경계 효과가 공간적 활성화 해석에 영향을 미칠 가능성을 보여준다. 따라서 dense 3D CNN 기반의 공간적 설명을 해석할 때에는 이러한 경계 효과를 함께 고려할 필요가 있다.

또한 ROAR는 특징 중요도 추정 기법이 실제로 모델 예측에 유효한 특징을 식별하는지를 검증하기 위해, 중요하다고 평가된 특징을 제거한 후 모델을 처음부터 재학습하여 성능 변화를 측정하는 벤치마크이다 [4]. 재학습 없이 특징 제거 후 재추론만 수행할 경우, 관찰된 성능 저하가 해당 특징의 정보 제거에 따른 것인지 또는 학습 분포와 다른 입력이 생성된 데 따른 것인지 명확히 구분하기 어렵다. 따라서 ROAR에서는 특징 제거 후 재학습 과정을 통해 이러한 영향을 최소화하고 특징 중요도 평가의 타당성을 검증한다.

따라서 본 논문에서는 GradientSHAP, Grad-CAM 및 ROAR를 함께 적용하여 3D

CNN 기반 의미론적 분할 모델의 예측 근거를 입력 채널과 공간적 활성화 영역의 두 측면에서 분석하고자 한다.

### 3. Dense 3D CNN 기반 XAI 분석 방법

#### 3.1 포인트 클라우드 데이터셋 및 전처리

본 논문에서는 항공사진측량 기반의 대규모 합성·실세계 포인트 클라우드 데이터셋인 STPLS3D를 활용하였다. STPLS3D는 합성 데이터인 Synthetic\_v1, Synthetic\_v2, Synthetic\_v3와 실세계 데이터인 Real World Data로 구성되어 있다 [6]. 본 연구에서는 XAI 기반 분석을 위해 Real World Data의 WMSC 타일을 대상으로 실험 영역을 구성하였다.

연산 자원의 제약을 고려하여 WMSC 타일 전체를 사용하지 않고, X, Y 평면상에서 40m × 40m 크기의 후보 영역을 크롭하여 사용하였다. 분석 대상 클래스가 충분히 포함된 영역을 확보하기 위해 Building과 Car 포인트가 각각 30개 이상 존재하는 것을 선정 조건으로 설정하고, 해당 조건을 만족하는 영역을 최대 40회 탐색하였다. 또한 학습 영역과 평가 영역은 공간적으로 중복되지 않도록 분리하였다.

선정된 크롭 영역으로부터 모델 입력 패치를 생성하였다. 각 패치는 X, Y, Z축 방향으로 각각 32개의 voxel로 구성하였으며, voxel 크기는 0.3m로 설정하였다. 이에 따라 하나의 패치는 각 축 방향으로 9.6m의 물리적 범위를 갖는다. Z축의 기준점은 각 크롭 영역의 최소 Z 좌표로 설정하였으며, 패치의 높이 범위를 벗어나는 포인트는 입력에서 제외하였다. 패치의 공간적 위치는 데이터셋 호출 시마다 새롭게 결정하여 동일한 크롭 영역 내에서도 다양한 위치의 패치가 학습에 활용되도록 하였다.

포인트 클라우드를 voxel로 변환한 후, 각

voxel의 semantic label은 내부에 포함된 point의 class를 기준으로 다수결(majority voting)을 통해 결정하였다. 두 개 이상의 class가 동일한 빈도를 갖는 경우에는 voxel 중심과 가장 가까운 point의 class를 해당 voxel의 label로 지정하였다. 또한 선행 연구와 동일하게 STPLS3D의 19개 세부 class를 Ground, Building, Tree, Car, Light pole, Fence의 6개 공통 class로 통합하였다. 구체적으로 Ground는 Ground, Road, Dirt, Grass의 4개 class, Building은 Building, Aircraft, Bike, Motorcycle, Clutter, Window의 6개 class, Tree는 Low vegetation, Medium vegetation, High vegetation의 3개 class, Car는 Vehicle, Truck, Military vehicle의 3개 class, Light pole은 Light pole과 Street sign의 2개 class로 통합하였으며, Fence는 기존 class를 그대로 사용하였다[1].

### 3.2 Full-Z, RGB-16 복셀화 기법

본 논문에서는 기존 연구에서 제시한 입력 표현을 기반으로 Baseline, RGB-16, Full-Z 및 Full-Z+RGB-16의 네 가지 복셀 표현을 구성하였다 [1]. Baseline은 voxel의 점유 여부를 나타내는 occupancy와 voxel 내부 point의 평균 RGB 값을 입력으로 사용하며, RGB-16은 occupancy와 색상 양자화를 통해 산출한 대표 RGB 값을 사용한다.

Full-Z는 각 voxel의 점유 정보와 높이 정보를 함께 표현하기 위해 occupancy와 normalized relative z-height의 2개 채널로 구성하였다. Occupancy는 해당 voxel 내부에 point가 존재하는지를 나타내며, normalized relative z-height는 크롭 영역을 기준으로 각 voxel의 상대적인 높이를 정규화한 값이다. 이를 통해 절대 좌표값에 대한 의존성을 줄이면서 객체의 수직적 위치 정보를 입력에 반영하였다. Full-Z 입력에는 RGB

색상 정보가 포함되지 않는다.

RGB-16은 voxel 내부 point의 R, G, B 값을 각각 16 단위로 양자화하여 색상 bin을 구성하고, 가장 많은 point가 포함된 bin을 해당 voxel의 지배적 색상 구간으로 선정한다. 이후 선정된 bin에 포함된 point의 원본 RGB 평균값을 해당 voxel의 대표 색상으로 사용하였다. 이러한 방식은 voxel 내부 모든 point의 RGB를 단순 평균하는 방식에 비해 지배적인 색상 특성을 보다 명확하게 보존하기 위한 것이다.

Full-Z+RGB-16은 Full-Z에서 사용하는 occupancy와 normalized relative z-height에 RGB-16을 통해 산출된 R16, G16, B16 채널을 결합한 입력 표현이다. 따라서 총 5개의 입력 채널로 구성되며, 본 연구의 XAI 분석에서는 occupancy, z-height, R16, G16, B16을 대상으로 각 채널의 기여도와 공간적 활성화 특성을 분석하였다.

### 3.3 Dense 3D CNN 학습 모델 구조

본 논문에서는 선행 연구에서 사용한 Sparse 3D CNN 대신, 설명가능성 기법, 특히 Grad-CAM을 표준적인 형태로 적용하기 위해 Dense 3D CNN 기반의 3D U-Net을 사용하였다. 입력 데이터의 표현 방식은 선행 연구의 정의를 유지하고, 모델의 연산 구조만 sparse 방식에서 dense 방식으로 변경하였다. 또한 본 연구에서는 제한된 크기의 patch를 입력으로 사용함으로써 Dense 구조에서 발생할 수 있는 메모리 사용량을 일정 범위 내에서 유지하였다.

Sparse 3D CNN은 전체 voxel grid 중 점유된 voxel을 중심으로 연산을 수행하는 sparse tensor 기반 구조로, 대규모 point cloud 처리 시 메모리 및 연산 효율이 높다는 장점이 있다. 그러나 sample마다 활성화되는 voxel의 위치와 개수가 달라 activation의 공간적 구조가 고정된

dense tensor와 동일하지 않으므로, 표준 Grad-CAM의 activation map과 gradient 추출 및 공간적 대응 과정을 그대로 적용하는 데 제약이 있다. 반면 Dense 3D CNN은 점유 여부와 관계없이 고정 크기의 voxel grid 전체를 tensor 형태로 처리하므로, convolution layer에서 일정한 공간 구조를 갖는 activation map과 이에 대응하는 gradient를 추출할 수 있다. 이에 따라 Grad-CAM을 이용한 voxel 공간상의 활성화 분석이 상대적으로 용이하다. 다만 비어 있는 voxel까지 연산에 포함하므로 voxel grid의 크기가 증가할수록 메모리 사용량과 연산량이 증가하는 한계가 있다.

본 연구에서는 STPLS3D의 일부 영역에서 생성한 32×32×32 voxel 규모의 patch를 모델 입력으로 사용하였다. 각 patch는 Full-Z+RGB-16 표현에 따라 occupancy, normalized relative z-height, R16, G16, B16의 총 5개 채널로 구성하였다.

모델은 encoder와 decoder로 구성된 3D U-Net 구조로 설계하였다. Encoder는 두 단계로 구성되며, 각 단계에서 3×3×3 Conv3D, BatchNorm3D 및 ReLU를 적용하였다. 특징 채널 수는 각각 16과 32로 설정하였으며, bottleneck에서는 64개의 특징 채널을 사용하였다. 각 encoder 단계의 공간 해상도 축소에는 MaxPool3D(2)를 적용하였다. Decoder에서는 stride 2의 transposed convolution을 이용하여 공간 해상도를 복원하고, 대응되는 encoder feature와 skip connection을 수행하였다. 이후 특징 채널 수를 각각 32와 16으로 축소하였으며, 최종적으로 1×1×1 Conv3D를 적용하여 6개 semantic class에 대한 voxel별 logit을 출력하였다.

모델 학습에는 클래스 불균형을 고려한 weighted Cross Entropy loss를 사용하였으며,

class가 할당되지 않은 voxel은 'ignore\_index = -1'로 설정하여 손실 계산에서 제외하였다. Optimizer는 Adam을 사용하였고 초기 learning rate는 0.001로 설정하였다. 기본 모델은 batch size = 4, epoch = 15 조건에서 학습하였으며, early stopping과 learning rate scheduler는 적용하지 않았다. ROAR 실험에서는 특징 제거 이후 모델을 재학습하였으며, 각 조건에서 80개의 patch를 사용하여 6 epoch 동안 학습하였다. 실험의 변동성을 확인하기 위해 random seed 0, 1, 2에 대해 반복 실험을 수행하였다.

### 3.4 Dense 3D CNN의 XAI 분석 방법

본 논문에서는 Dense 3D CNN의 예측 근거를 분석하기 위해 Grad-CAM, GradientSHAP 및 ROAR를 적용하였다.

Grad-CAM은 마지막 convolution block에서 추출한 activation과 gradient를 이용하여 목표 클래스에 대한 3차원 activation map을 생성한다. 생성된 activation map은 occupancy voxel과 함께 시각화하여 모델이 목표 클래스를 예측하는 과정에서 공간적으로 어떠한 영역에 반응하는지를 정성적으로 분석하였다.

GradientSHAP은 입력 voxel에 대한 attribution을 계산하고, 이를 occupancy, z-height, R16, G16, B16의 각 입력 채널별 평균 절댓값으로 요약하여 클래스별 채널 기여도를 비교하였다. 본 실험에서는 목표 클래스가 포함된 patch 5개를 확보하는 것을 기준으로 최대 200회까지 탐색하였다. 이때 n\_samples=20은 분석에 사용되는 patch의 수가 아니라, 각 patch에 대한 GradientSHAP attribution 계산에 사용되는 샘플 수를 의미한다. 목표 출력은 해당 patch 전체에서 목표 클래스에 대한 logit의 평균값으로 설정하였으며, 기준 입력(baseline)은 모든 값이 0인

tensor를 사용하였다.

ROAR 실험에서는 SHAP 기반 중요도 순위에 따라 입력 채널을 순차적으로 제거한 후 모델을 처음부터 재학습하고, 이에 따른 Car 클래스 IoU의 변화를 측정하였다. Full-Z+RGB-16 입력은 총 5개의 채널로 구성되므로 20%, 40%, 60%, 80%, 100%의 제거율은 각각 1개, 2개, 3개, 4개, 5개 채널의 제거에 해당한다. 각 제거 조건에 대해 seed 0, 1, 2를 적용하여 반복 실험을 수행하였으며, 동일한 제거율에서 무작위로 채널을 제거한 결과와 비교하여 SHAP 기반 중요도 순위의 유효성을 검증하였다.

#### 4. Dense 3D CNN의 XAI 분석 결과

##### 4.1 SHAP 기반 입력 채널 중요도 분석

Full-Z+RGB-16 복셀 표현으로 학습된 Dense 3D CNN을 대상으로 Building과 Car 클래스에 대한 GradientSHAP 분석을 수행하였다. 그림 1과 그림 2는 occupancy, z-height, R16, G16, B16 각 입력 채널의 평균 절대 attribution을 나타낸다.

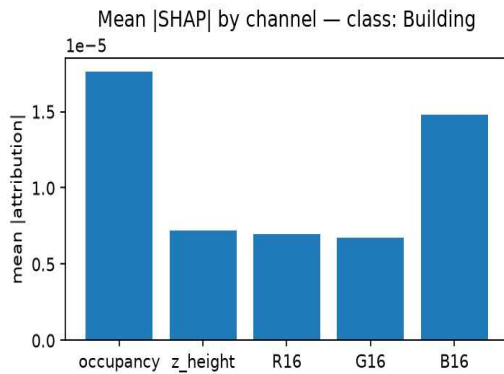


그림 1. Building 클래스에 대한 채널별 평균 |SHAP| 값  
Fig. 1. Mean |SHAP| values by input channel for the Building class

분석 결과, Building과 Car 클래스 모두에서

occupancy 채널의 평균 절대 attribution이 가장 높게 나타났다. 이는 본 실험에 사용된 모델과 분석 표본에서 occupancy 정보가 Building과 Car 클래스의 예측에 상대적으로 큰 영향을 미쳤음을 의미한다. 다만 이러한 결과는 본 연구에서 설정한 실험 조건에 기반한 것으로, 모든 STPLS3D 영역이나 객체 클래스에서 형상 정보가 색상 및 높이 정보보다 항상 높은 중요도를 갖는다는 의미로 해석하지 않았다.

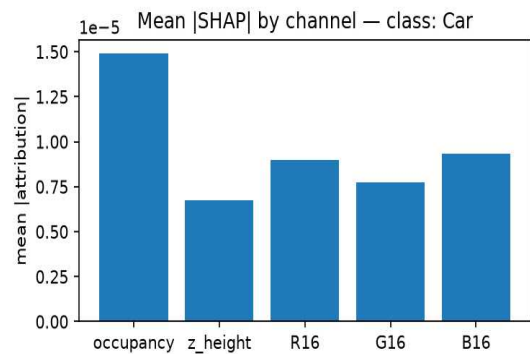


그림 2. Car 클래스에 대한 채널별 평균 |SHAP| 값  
Fig. 2. Mean |SHAP| values by input channel for the Car class

Building 클래스에서는 occupancy 다음으로 B16의 기여도가 높게 나타났으며, 이후 z-height, R16, G16 순으로 나타났다. Car 클래스에서도 occupancy가 가장 높은 기여도를 보였으며, B16과 R16이 그 다음으로 높게 나타난 반면, z-height는 분석된 5개 채널 중 가장 낮은 기여도를 보였다.

이러한 결과는 입력 채널의 상대적 중요도가 목표 클래스에 따라 다르게 나타날 수 있음을 보여준다. 특히 Car 클래스에서 z-height의 기여도가 상대적으로 낮게 나타났으나, 이를 높이 정보 자체의 유용성이 낮다는 의미로 해석하기에는 한계가 있다. GradientSHAP에 의해 산출되는 기여

도는 분석에 사용된 크롭 영역과 patch 표본, 학습된 모델의 파라미터 및 기준 입력 등의 영향을 받을 수 있으므로, 보다 다양한 공간 영역과 반복 실험을 통한 추가적인 검증이 필요하다.

#### 4.2 Grad-CAM 기반 공간적 활성화 분석

그림 3은 Car 클래스에 대한 입력 occupancy와 Grad-CAM heatmap을 비교한 결과이다. Grad-CAM 분석 결과, 실제 점유 voxel 주변에서 activation이 나타나는 동시에 입력 patch의 일부 경계에서도 상대적으로 높은 activation이 나타났다.

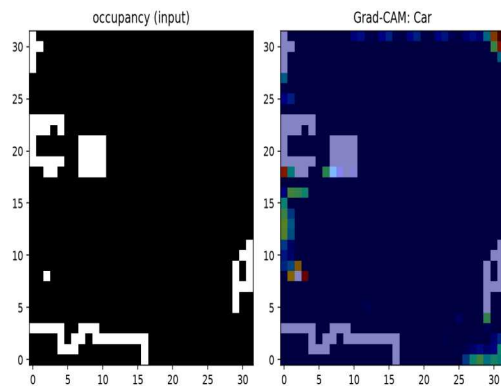


그림 3. Car 클래스에 대한 Grad-CAM 시각화  
(좌: 입력 occupancy, 우: Grad-CAM heatmap)  
Fig. 3. Grad-CAM visualization for the Car class  
(left: input occupancy; right: Grad-CAM heatmap)

특히 patch 경계와 일치하는 위치에서 나타난 일부 activation은 실제 객체의 occupancy 분포와 직접적으로 대응하지 않는 것으로 확인되었다. 이러한 현상은 Dense 3D CNN의 convolution 및 padding 과정에서 발생하는 경계 효과와 관련될 가능성이 있다. 다만 본 실험 결과만으로 zero-padding이 해당 activation의 직접적인 원인이라고 단정하기는 난해하다. 따라서 patch 경계에서 나타나는 Grad-CAM activation을 객체의 의미적 특징으로 직접 해석하는 데에

는 주의가 필요하다. 본 연구에서는 Grad-CAM 결과를 occupancy 분포와 함께 비교하여 해석하였으며, 객체의 occupancy와 직접적으로 대응하지 않고 경계 부근에서 나타나는 activation은 보조적인 현상으로 고려하였다. 향후에는 overlapping patch의 적용, 경계 영역 제외 및 padding 방식 변경 등을 통해 경계 activation의 발생 원인을 추가적으로 검증할 필요가 있다.

#### 4.3 ROAR 기반 3D CNN 학습 모델의 정량적 검증

그림 4는 SHAP 중요도 순위와 무작위 순위에 따라 입력 채널을 제거한 후 모델을 처음부터 재학습하여 측정한 Car 클래스 IoU의 변화를 나타낸다.

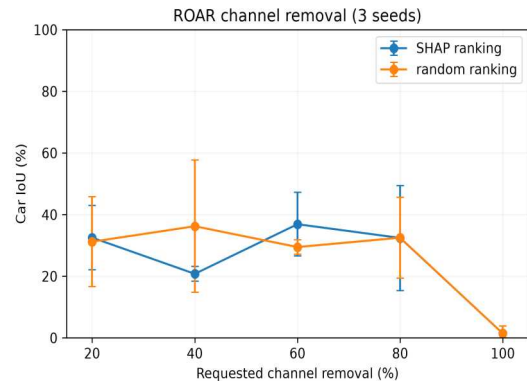


그림 4. 입력 채널 제거율에 따른 Car IoU(3개 seed의 평균 ± 표본 표준편차)

Fig. 4. Car IoU at actual channel removal rates (mean ± sample standard deviation over three seeds)

Full-Z+RGB-16은 occupancy, z-height, R16, G16, B16의 총 5개 입력 채널로 구성되므로, 20%, 40%, 60%, 80%, 100%의 채널 제거율은 각각 1개, 2개, 3개, 4개, 5개 채널의 제거에 해당한다. 각 조건은 seed 0, 1, 2에서 반복하였으며, 결과는 세 seed에서 산출된 IoU의 평균과 표본 표준편차로 제시하였다.

SHAP 중요도 순위에 따른 채널 제거 결과, Car IoU는 제거율 20%, 40%, 60%, 80%, 100%에서 각각  $32.55 \pm 10.40\%$ ,  $20.79 \pm 2.42\%$ ,  $36.89 \pm 10.32\%$ ,  $32.43 \pm 17.04\%$ ,  $1.43 \pm 2.42\%$ 로 나타났다. 무작위 채널 제거에서는 각각  $31.20 \pm 14.56\%$ ,  $36.24 \pm 21.45\%$ ,  $29.48 \pm 2.35\%$ ,  $32.52 \pm 13.09\%$ ,  $1.43 \pm 2.42\%$ 로 나타났다.

SHAP 중요도 순위에 따른 제거에서는 20% 조건에서 occupancy가 우선 제거되었으며, 40%에서는 occupancy와 B16, 60%에서는 occupancy, R16, B16, 80%에서는 occupancy, R16, G16, B16이 제거되었다. 100% 조건에서는 5개의 입력 채널이 모두 제거되었다.

특히 40% 제거 조건에서 SHAP 중요도 순위에 따른 제거 후 Car IoU는 20.79%로 나타났으며, 무작위 제거 결과인 36.24%보다 15.45%p 낮았다. 이는 해당 조건에서 SHAP이 상대적으로 높은 중요도를 부여한 채널을 제거하였을 때 모델의 성능 저하가 더 크게 나타났음을 보여주는 보조적 근거이다.

그러나 채널 제거율이 증가함에 따라 IoU가 단조롭게 감소하는 경향은 나타나지 않았다. 예를 들어 SHAP 중요도 순위에 따른 제거에서는 40% 조건에서 Car IoU가 20.79%까지 감소한 이후 60% 조건에서 36.89%로 다시 증가하였다. 또한 무작위 제거에서는 seed에 따라 제거되는 채널 조합이 달라 일부 조건에서 비교적 큰 표준편차가 나타났다. 이러한 변동은 입력 채널 간 상호작용뿐만 아니라 제한된 patch 수, 재학습 epoch 및 seed에 따른 영향을 포함할 가능성이 있다.

따라서 본 ROAR 실험 결과는 SHAP 중요도 순위의 절대적인 타당성을 입증하기보다는, 중요도가 높게 평가된 채널의 제거에 따른 모델 성능 변화를 정량적으로 확인하기 위한 보조적 검증 결과로 해석하였다. 특히 본 실험은 5개의 입력

채널과 3개의 seed를 기반으로 수행되어 제거율을 세밀하게 설정하고 반복 실험의 변동성을 충분히 평가하는 데 한계가 있다. 향후에는 보다 많은 seed와 patch 및 충분한 재학습 epoch를 적용하여 결과의 재현성을 추가적으로 검증할 필요가 있다.

## 5. 결 론

본 논문에서는 occupancy, normalized relative z-height, R16, G16, B16의 5개 채널로 구성된 Full-Z+RGB-16 입력을 사용하는 Dense 3D CNN을 대상으로 GradientSHAP, Grad-CAM 및 ROAR를 적용하여 모델의 예측 근거를 입력 채널과 공간적 활성화의 측면에서 분석하였다.

GradientSHAP 분석 결과, 본 실험에서 분석한 Building과 Car 클래스 모두 occupancy 채널의 평균 절대 attribution이 가장 높게 나타났다. 반면 z-height와 RGB 채널의 상대적 중요도는 클래스에 따라 서로 다른 양상을 보였다. 이러한 결과는 분석에 사용된 크롭 영역과 patch 및 모델 조건에 기반한 것으로, 해석 범위를 본 연구의 실험 조건으로 한정하였다.

Grad-CAM 분석에서는 실제 점유 voxel 주변뿐만 아니라 patch 경계에서도 activation이 관찰되었다. 일부 경계 activation은 occupancy 분포와 직접적으로 대응하지 않았으며, Dense 3D CNN의 convolution 과정에서 발생하는 경계 효과와 관련될 가능성을 확인하였다. 다만 본 실험만으로 zero-padding과 경계 activation 간의 직접적인 인과관계를 규명하기에는 한계가 있다. 따라서 Grad-CAM을 voxel 기반 의미론적 분할 모델에 적용할 경우 객체 영역의 activation뿐만 아니라 patch 경계에서 발생하는 activation을 함께 고려하여 해석할 필요가 있다.

ROAR 실험에서는 40% 채널 제거 조건에서

SHAP 중요도 순위에 따른 제거 후 Car IoU가 20.79%로 나타났으며, 무작위 제거 결과인 36.24%보다 15.45%p 낮게 나타났다. 이는 해당 조건에서 SHAP이 높은 중요도를 부여한 채널의 제거가 모델 성능에 상대적으로 큰 영향을 미쳤음을 보여주는 보조적 근거이다. 그러나 제거율 증가에 따른 IoU의 변화가 단조 감소 형태를 보이지 않았으며, 일부 조건에서는 seed 간 높은 변동성도 관찰되었다.

따라서 본 연구의 XAI 분석 결과는 제한된 크롭 영역, 모델 설정, 분석 patch 및 3개의 seed를 기반으로 수행된 실험 결과로서 해당 실험 조건의 범위에서 해석할 필요가 있다. 향후에는 STPLS3D의 여러 독립적인 타일을 대상으로 seed 수, 학습 epoch 및 patch 수를 확대하여 SHAP 기반 채널 중요도와 ROAR 결과의 재현성을 검증할 필요가 있다. 또한 overlapping patch와 다양한 padding 조건을 적용한 Grad-CAM 분석을 통해 patch 경계에서 발생하는 activation의 원인을 추가적으로 분석하고자 한다.

본 연구는 국토교통부와 국토교통과학기술진흥원의 지원으로 수행되었음 (과제번호 : RS-2025-02317649, 국토정보 고도화를 위한 입체격자체계 적용 및 활용 기술 개발)

## 참 고 문 헌

- [1] J. Kim, K. Park, J. Seo, K. Lee, K. Kim, and Y. Lee, "Performance Comparison of Voxelization Methods for 3D CNN-Based Urban 3D Data Semantic Segmentation," in Proc. The 7th Korea Artificial Intelligence Conference, 2026.  
URL: <https://koreaai.org/proceeding>
- [2] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Conference on Neural Information Processing Systems (NeurIPS), pp. 4765-4774, 2017.  
DOI: 10.48550/arXiv.1705.07874
- [3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017.  
DOI: 10.1109/ICCV.2017.74
- [4] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A Benchmark for Interpretability Methods in Deep Neural Networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.  
DOI: 10.48550/arXiv.1806.10758
- [5] A. Kuriyal and V. Kumar, "Towards Explainable LiDAR Point Cloud Semantic Segmentation via Gradient Based Target Localization," arXiv preprint arXiv:2402.12098, 2024.  
DOI: 10.48550/arXiv.2402.12098
- [6] M. Chen, Q. Hu, Z. Yu, H. Thomas, A. Feng, Y. Hou, K. McCullough, F. Ren, and L. Soibelman, "STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset," in Proc. British Machine Vision Conference (BMVC), 2022.  
<https://bmvc2022.mpi-inf.mpg.de/0429.pdf>
- [7] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep Learning for 3D Point Clouds: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 12, pp. 4338-4364, 2021.  
DOI: 10.1109/TPAMI.2020.3005434
- [8] A. Xiao, J. Huang, D. Guan, X. Zhang, S. Lu, and L. Shao, "Unsupervised Point Cloud Representation Learning With Deep Neural Networks: A Survey," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 11321–11339, 2023.

DOI: 10.1109/TPAMI.2023.3262786

- [9] M. De Vries, R. Naidoo, O. Fourkioni, L. G. Dent, N. Curry, C. Dunsby, and C. Bakal, “Interpretable Point Cloud Classification Using Multiple Instance Learning,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22209–22220, 2025.

DOI: 10.1109/ICCV51701.2025.02062

- [10] R. R. Selvaraju, P. Chattopadhyay, M. Elhoseiny, T. Sharma, D. Batra, D. Parikh, and S. Lee, “Choose Your Neuron: Incorporating Domain Knowledge Through Neuron-Importance,” in Proc. European Conference on Computer Vision (ECCV), pp. 526–541, 2018.

DOI: 10.1007/978-3-030-01261-8\_32

- [11] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, “C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11676–11685, 2022.

DOI: 10.1109/CVPR52688.2022.01138

#### 저 자 소 개



김재홍(Jaehong Kim)

2021.3 동국대학교 컴퓨터AI학부 입학  
<주관심분야> 공간정보, 분산 AI, 컴퓨터 비전



서재석(Jaeseok Seo)

2021. 3 건양대학교 의료인공지능학과 입학  
<주관심분야> 컴퓨터 비전, 객체 탐지, 의료 AI



박기욱(Kiwook Park)

2021.3 울산대학교 IT융합학과 입학  
<주관심분야> 백엔드 시스템, 클라우드 인프라, AX



김광수(Kwangsoo Kim)

1993.2 고려대학교 정보공학과 졸업  
1995.2 고려대학교 전산학과 석사  
2016.8 충남대학교 컴퓨터공학과 박사  
1994.12-현재 : 한국전자통신연구원 책임연구원  
<주관심분야> 공간정보, 디지털 트윈, 사물인터넷, 위치기반 시스템, 공간 데이터 큐브



이기훈(Kihoon Lee)

2018.2 청주대학교 지적학과 졸업  
2020.2 청주대학교 지적학과 석사  
2024.02 청주대학교 도시·부동산·지적학 박사  
2026.1~ 현재 : 한국전자통신연구원 박사후연구원  
<주관심분야> AI 기반 스마트 시티, GeoAI,  
자연어 처리, 디지털 트윈, 공간정보