THE JOURNAL OF KOREAN INSTITUTE OF ELECTROMAGNETIC ENGINEERING AND SCIENCE. 2026 October; 37(10), 1055～1064.

https://doi.org/10.5515/KJKIEES.2026.37.10.1055
ISSN 1226-3133 (Print) · ISSN 2288-226X (Online)

확산 모델 기반 레이다-카메라 융합 깊이 추정에 관한 연구

Radar-Camera Fusion Depth Estimation Based on Diffusion Model

이재룡 · 신정완* · 신헌용** · 김동영** · 김민*** · 최재호*

Jae-Ryong Lee · Jeong-Wan Shin* · Hun-Yong Shin** · Dong-Young Kim** · Min Kim*** · Jae-Ho Choi*

요 약

본 논문은 자율주행 환경에서 사전 학습된 대규모 확산 모델의 풍부한 기하학적 사전 지식을 레이다-카메라 융합 깊이 추정 작업에 그대로 적용할 때 발생하는 모달리티 불일치, 노이즈 스케줄, 모델 용량의 문제를 분석하고 해결책을 제시한다. 첫째, 확산을 통한 생성 과정 대신 판별 모델 기반의 단일 패스 회귀 구조를 도입하고 픽셀 공간의 직접적인 손실 함수를 적용한다. 둘째, 기존 텍스트 어텐션을 재사용하는 대신 무작위 초기화된 신규 크로스 어텐션 블록을 구축하여 언어 도메인 편향으로 인한 오류를 제거한다. 셋째, 거대한 백본이 레이다 경로를 무시하고 이미지 경로만 신뢰하는 현상을 방지하기 위해 U-Net을 완전 동결하고 좁은 하위 네트워크만 학습하도록 제한한다. 대규모 nuScenes 데이터셋을 활용한 실험 결과, 제안 방법론은 절대 깊이 스케일 모호성을 성공적으로 해결하며 평균 절대 오차(MAE) 3.72 m를 달성하여 단일 카메라 베이스라인 대비 성능을 크게 개선한다.

Abstract

This study introduces three deep learning architectural design choices to effectively adapt pretrained diffusion models for radar-camera fusion depth estimation. First, we replace the iterative reverse diffusion process with a predictive single-pass regression framework, enforcing direct pixel-space loss. Second, rather than hijacking existing text attention layers, we implement randomly initialized fresh cross-attention blocks. Third, we freeze the U-Net core and restrict training to a small subset of parameters to prevent the large backbone from bypassing the radar pathway by exploiting image-only shortcuts. Evaluated on the nuScenes dataset, our approach successfully eliminates metric scale ambiguity and significantly improves accuracy, achieving a mean absolute error (MAE) of 3.72 m, compared with the monocular diffusion baseline.

Key words: Deep Learning, Diffusion, Radar-Camera Fusion, Depth Estimation

Ⅰ. 서 론

카메라는 3차원의 현실 세계를 2차원의 이미지로 투영하여 기록한다. 이 과정에서 각 픽셀까지의 깊이 정보가 손실되는데, 이를 복원하는 것은 증강현실, 자율주행 및 로봇 공학 등 수많은 산업의 필수 과제이다. 단일 카메라

「이 연구는 2025년 정부(방위사업청)의 재원으로 국방과학연구소의 지원을 받아 수행된 미래도전국방기술 연구개발사업임(No. 915141301).」

대구경북과학기술원 기초학부(School of Undergraduate Studies, Daegu Gyeongbuk Institute of Science & Technology)

*대구경북과학기술원 전기전자컴퓨터공학과(Department of Electrical Engineering & Computer Science, Daegu Gyeongbuk Institute of Science & Technology)

**국방과학연구소(Agency of Defence Development)

***국립부경대학교 스마트모빌리티공학과(Department of Smart Mobility Engineering, Pukyong National University)

· Manuscript received June 29, 2026 ; Revised July 9, 2026 ; Accepted July 30, 2026. (ID No. 20260629-061)

· Corresponding Author: Jae-Ho Choi (e-mail: jhochoi@dgist.ac.kr)

ⓒ Copyright The Korean Institute of Electromagnetic Engineering and Science. This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

이미지로부터 3차원 깊이 정보를 추정하는 모델은 딥러닝을 활용하여 이미지 속 물체의 크기, 음영, 질감 등을 학습하여 깊이를 예측한다. 초기에는 정교한 판별 모델을 사용해왔지만, 최근에는 Stable Diffusion[1] 기반 모델인 Marigold[2]와 같이 대규모 데이터셋으로 사전 학습된 확산 모델의 잠재 공간에 내재된 자연계의 기하학적 사전 지식을 활용하여 제로 샷(zero-shot) 환경에서도 미세한 디테일과 물체의 경계를 복원해 내는 생성형 깊이 추정 모델이 등장하고 있다.

하지만 단일 이미지만으로는 정확한 절대 깊이를 추정하기 힘들고, 특히 어두운 환경이나 빛 번짐, 악천후 환경에서의 깊이를 추정하는 데에 어려움이 있다. 이를 극복하기 위해 레이다 센서를 결합한 레이다-카메라 융합 연구도 활발히 진행 중이다. 레이다는 거리에 구애받지 않는 거리 해상도를 가져 탐지 대상과의 거리가 멀어져도 보다 정확한 측정이 가능하다[3]. 또한 레이다는 라이다 대비 압도적인 비용 효율성을 보여주며, 악천후와 같은 열악한 환경에서도 안정적인 성능을 보여준다. 하지만 레이다의 극심한 희소성과 수직 해상도 정보의 부재 및 노이즈는 레이다 데이터를 2D 이미지 평면에 단순히 사영하는 것만으로는 깊이를 추정하기 어렵게 만든다. 이러한 문제를 해결하기 위해 풍부한 구조적 정보가 담겨 있는 카메라 이미지와의 융합을 통해 두 모달리티의 장단점을 상호보완적으로 활용하여 깊이를 추정하는 연구가 진행 중이다.

따라서 레이다-카메라 융합 연구들은 레이다의 공간적 공백을 메우고 두 모달리티의 특징을 정렬하기 위해 다양한 기법을 사용한다. 참고문헌 [4]~참고문헌 [6]에서는 두 단계로 모델을 구성하여 희소한 레이다 포인트를 중간 단계 깊이 맵으로 퍼뜨려 밀도를 높인 다음 다시 이미지 특징과 융합하여 최종 깊이 맵을 추정한다. 이후 참고문헌 [7]에서는 레이다의 기하학적 정보 추출을 위한 GNN 인코더와 계층적 융합을 활용해 효율적인 깊이 추정을 가능하게 한다.

본 연구에서는 레이다의 희소성과 노이즈 문제를 해결하기 위해 최근 단일 카메라 깊이 추정에 활용되고 있는 Stable Diffusion 모델을 사용하고자 한다. Stable Diffusion과 같은 대규모 확산 모델은 방대한 이미지 데이터셋으로 학습되어 현실 세계의 사실적이고 세밀한 구조적 정보를 복원하는 능력이 탁월하다. 이에 본 연구에서는 확산 모델의 풍부한 사전 지식을 활용함으로써, 물리적 센서의 한계로 인해 발생하는 레이다의 노이즈를 효과적으로 교정하고 현실 세계의 기하학적 세부 구조를 정밀하게 복원하고자 한다.

하지만 본질적으로 Stable Diffusion 모델은 텍스트 조건부 생성 모델로 설계되었기에, 이를 정밀한 회귀 작업인 깊이 추정에 그대로 이식하는 과정에서 세 가지 문제가 발생한다. 첫째, 텍스트와 성격이 완전히 다른 레이다 특징을 주입하는 과정에서의 모달리티 불일치가 생긴다. 둘째, 생성 이미지의 품질을 높이기 위한 확률에 기반한 노이즈 스케줄이 결정론적 깊이 추정에 오히려 방해가 될 수 있다. 셋째, 거대한 모델의 표현 용량이 불완전한 레이다 정보를 압도하거나 무시해 버리는 문제가 발생할 수 있다.

본 논문에서는 사전 학습된 대규모 확산 모델을 레이다-카메라 깊이 추정에 적용할 때 발생하는 병목 현상을 분석하고, 이를 해결하기 위한 방법을 제시한다. 깊이 추정에 확산 기반 학습보다 판별 모델 방식이 유리함을 보이고, Stable Diffusion의 기존 텍스트 크로스 어텐션 블록을 재사용하는 것의 문제를 입증한다. 또 U-Net 전체를 파인튜닝하는 방식보다 가중치를 동결하고 좁은 하위 네트워크만 학습하는 방식의 성능이 우수함을 보이고자 한다.

Ⅱ. 모델 파이프라인

2-1 모델의 구조

그림 1은 모델의 전체적인 학습 과정을 그린 구조도이다. 전체 파이프라인의 입력은 RGB 이미지와 레이다 포인트 클라우드이다. 전체 추정 프로세스는 사전 학습된 Stable Diffusion 모델을 바탕으로 한다. 먼저 입력 이미지는 VAE 인코더를 통해 잠재 공간으로 인코딩되고, 레이다 포인트 클라우드는 공간적 맥락을 추출하기 위해 TacoDepth에서 제안한 GNN 인코더를 사용한다. 이때, 레이다 포인트의 개수는 매 프레임마다 다르므로 한 배치로 묶기 위해 zero-padding 또는 랜덤 샘플링을 통해 128

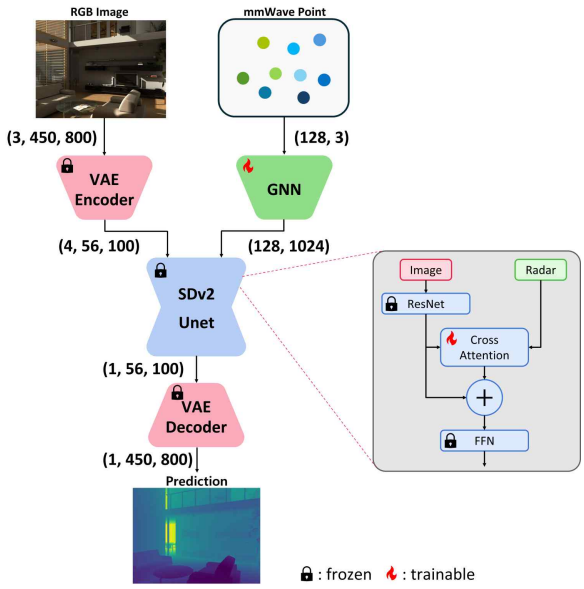


그림 1. 모델의 구조
Fig. 1. Model architecture.

개로 맞춰준다. Zero-padding된 포인트가 실제 레이다로 취급되지 않도록 이진 마스크를 함께 모델에 입력한다. GNN 인코더에서는 먼저 각 레이다 포인트에 대해 주변 포인트로부터 노드 특징을 추출하고, 이를 바탕으로 레이다 포인트들을 그래프로 구성한다. 이를 통해 자세한 위치 정보와 구조적 특징을 추출하게 된다.

이렇게 인코딩된 레이다 특징은 U-Net의 각 레이어의 해상도 단계에 맞춰 개별적으로 크로스 어텐션 블록에 주입된다. 각 크로스 어텐션 블록 내에서 쿼리(query)는 U-Net 내부의 이미지 잠재 특징으로부터 생성되며, 키(key)와 값(value)은 주입된 레이다 특징들로부터 생성된다. 이를 통해 이미지의 각 픽셀이 자신과 기하학적으로 연관된 레이다 신호를 참조할 수 있게 된다. 여기서 레이다 특유의 수직 해상도 정보의 모호성을 고려하여, 키와 값으로 선택하는 레이다 포인트는 픽셀 기준 상하 거리에는 제한을 두지 않고 좌우 거리에 제한을 둔 범위 안의 포인트들을 선택하여 어텐션 연산을 수행한다. 이후 U-Net을 통해 융합된 특징은 VAE 디코더를 거쳐 최종 깊이로 변환되게 된다.

2-2 판별 모델 기반 단일 패스 추정

대규모 데이터로 사전 학습된 Stable Diffusion 모델은 이미지의 확률 분포를 학습하여 노이즈로부터 이미지를 생성하므로 확률적 프로세스가 필요하다. 반면, 깊이 추정은 확실한 정답이 있는 회귀 문제이므로 이 과정에서 점진적으로 노이즈를 주입하고 제거하는 확산 모델의 노이즈 스케줄은 오히려 레이다 신호의 정보를 희석하고 학습을 방해하는 노이즈로 작용한다.

본 연구에서는 확산 기반의 다단계 생성 프로세스를 포기하고, U-Net이 단 한 번의 순방향 연산으로 깊이를 회귀하는 판별 모델 기반의 단일 패스 방식을 채택한다. 기존의 확산 모델이 잠재 공간에서의 노이즈 예측에 대한 L2 손실 함수를 사용하는 것과 달리, U-Net이 입력 이미지와 레이다 조건을 받아 최종 깊이의 잠재 표현을 즉시 출력하며, VAE 디코딩 후 픽셀 공간에서 정답과의 L1 손실 함수를 통해 직접 학습된다. 정답으로는 참고문헌 [4]에서 사용한 데이터 전처리 방식을 따라 단일 라이다 프레임 d_{gt}와, 단일 라이다 프레임의 희소성을 보완하여 사물의 형태를 대략적으로 복원하기 위해 앞뒤 80 프레임을 누적한 뒤 스캐폴딩[8] 방식을 사용하여 보간한 깊이 d_{acc}를 사용한다.

$$L = \frac{1}{|\Omega_{gt}|}\sum_{x \in \Omega_{gt}} \left|\hat{d}(x) - d_{gt}(x)\right| + \frac{1}{|\Omega_{acc}|}\sum_{x \in \Omega_{acc}} \left|\hat{d}(x) - d_{acc}(x)\right| \quad (1)$$

Ω_{gt}와 Ω_{acc}는 각각 단일 깊이와 보간한 깊이가 유효한 값을 가지고 있는 픽셀, $\hat{d}$는 모델이 예측한 깊이를 의미한다. 첫째 항은 단일 라이다 프레임 d_{gt}에 대한 L1 손실, 둘째 항은 보간한 깊이 d_{acc}에 대한 L1 손실이 되어 절대적인 거리를 추정함과 동시에 대략적인 기하 구조를 반영할 수 있도록 한다.

2-3 기존 텍스트 크로스 어텐션 블록 초기화

Stable Diffusion 모델은 본래 텍스트 조건부 이미지 생성 모델이므로 텍스트 조건에 맞는 어텐션 블록의 가중

치를 갖고 있다. 하지만 본 연구의 경우 레이다를 조건으로 사용하여 구조적인 모달리티 불일치를 야기한다. 사전 학습된 어텐션 블록은 추상적인 언어 공간의 분포에 정렬되어 있다. 여기에 완전히 다른 특성과 차원을 가진 레이다 특징을 주입하면, 모델은 비정상적인 출력을 내거나 학습 과정에서 레이다 신호의 어텐션 가중치가 소실되는 현상이 발생하게 된다. 따라서 기존 어텐션 블록의 가중치를 무시하고, 레이다 특징 공간에 맞게 어텐션 블록을 새롭게 정의하여 레이다 조건 주입을 위해 사용한다.

2-4 U-Net 동결 및 좁은 하위 모델 파라미터 학습

대규모 백본 모델의 전체 네트워크를 파인튜닝하는 것으로 다른 하위 작업에 적용시키는 기법은 여러 문제에 적용되고 있는 일반적인 기법이다. 하지만 레이다-카메라 깊이 추정에서는 적합하지 않다. 표 1에서 볼 수 있듯이, U-Net의 학습 가능한 파라미터는 레이다 정보가 들어가는 GNN 인코더와 크로스 어텐션 블록에 비해 매우 큰 부분을 차지하고 있다. 만약 U-Net 전체를 학습 가능하게 열어두면, 모델은 입력된 이미지로부터 깊이 정보를 추출하는 지름길 경로를 너무 빠르게 학습해 버린다. 즉, 거대한 표현 용량을 가진 이미지 백본이 우세하게 작동하면서, 학습이 되어있지 않고 상대적으로 희소한 레이다 경로로 흘러들어가야 할 정보를 고갈시키는 현상이 발생한다.

따라서 본 논문에서는 대규모 이미지 백본의 가중치를 동결함으로써 이미지 특징의 영향을 막고, 오직 레이다 조건 주입과 관련된 좁은 하위 네트워크만을 학습하도록 구현한다. 이 파라미터 용량의 제한은 네트워크로 하여금 깊이를 정확하게 예측하기 위해 레이다 경로의 정보유입을 강제로 선택하고 의존하도록 만드는 역할을 수행한다.

표 1. 각 모듈별 파라미터 크기
Table 1. Size of parameters for each module.

Module	Parameters (M)
VAE	83.7
U-Net	870
Cross-attention blocks	50.4
GNN encoder	2.49

Ⅲ. 실험 설계 및 결과

3-1 실험 설계 상세

본 연구는 대규모 자율주행 데이터셋인 nuScenes[9]를 사용하여 모델을 학습하고 평가한다. nuScenes 데이터셋은 복잡한 도심 환경에서 수집된 카메라 이미지와 레이다 포인트 클라우드, 그리고 라이다 포인트 클라우드를 제공한다. 보스턴과 싱가포르에서 수집한 총 1000개의 도로 주행 장면이 있으며, 공식 데이터셋 분할 프로토콜에 따라 학습에 700장면, 검증과 평가에 각 150개의 장면을 사용한다.

nuScenes 데이터셋은 Continental 사의 ARS 408-21 레이다를 사용한다. 이 센서는 77 GHz 주파수 대역에서 동작하는 FMCW 레이다로 13 Hz의 데이터 수집 주기를 갖는다. 레이다 데이터는 nuScenes에서 각 프레임마다 N개의 포인트의 3D 위치, RCS, Doppler 등 18개 특성의 정보가 담긴 [18, N] 배열로 제공된다. 이를 참고문헌 [4]를 따라 nuScenes에서 제공한 레이다와 카메라 캘리브레이션 및 차량 움직임 정보를 바탕으로 레이다 포인트를 카메라 좌표계로 옮겨 카메라 뒤쪽 포인트와 이미지 밖으로 투영되는 포인트를 제거한 뒤, 그 포인트들의 3D 위치가 담긴 [N, 3] 배열로 반환하여 모델에 입력한다.

백본 모델은 Stable Diffusion v2 모델을 사용하고, 원래 텍스트 임베딩을 받던 Stable Diffusion v2 U-Net의 각 레이어에 존재하는 크로스 어텐션 블록을 초기화하여 레이다 임베딩을 주입한다. 학습 시에는 크로스 어텐션 블록을 제외한 U-Net과 VAE 인코더 및 디코더를 동결하고 크로스 어텐션 블록과 레이다 임베딩을 위한 GNN 인코더만을 학습한다. Optimizer는 AdamW를 사용하고, 학습률은 $\eta_{\max}$=3.0e-5에서 $\eta_{\min}$=3.0e-7로 감쇠하는 Cosine Annealing LR 스케줄러를 사용한다. 학습 환경은 NVIDIA RTX PRO 6000 Blackwell GPU에서 배치 크기 16으로 4000 iterations 동안 학습한다.

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{2}$$

학습 중에는 기준 단일 라이다 프레임과 함께, 앞뒤 80프레임을 누적한 뒤 스캐폴딩 방식을 사용하여 보간한 깊이를 정답으로 사용한다. 평가 시에는 단일 라이다 프

레임을 사용하고, 성능 분석을 위해 기존 레이다-카메라 융합 깊이 추정 연구[4]-[7]를 따라 0∼50 m, 0∼70 m, 0∼80 m에서의 평균 절대 오차(MAE)와 평균 제곱근 오차(RMSE)를 바탕으로 모델을 평가한다.

$$MAE = \frac{1}{|\Omega_{gt}|} \sum_{x \in \Omega_{gt}} \left| \hat{d}(x) - d_{gt}(x) \right| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{|\Omega_{gt}|} \sum_{x \in \Omega_{gt}} \left| \hat{d}(x) - d_{gt}(x) \right|^2} \tag{4}$$

단일 이미지 기반의 제로 샷 깊이 추정에서 Stable Diffusion 모델을 활용하여 최첨단 성능을 기록한 Marigold를 주요 베이스라인으로 설정한다. 사전 학습된 확산 모델의 기하학적 잠재력이 희소한 레이다 조건과 결합됨으로써 생기는 절대 깊이 추정 성능의 변화를 nuScenes로 학습한 Marigold 모델과의 비교를 통해 명확히 보이고자 한다. 또한 기존의 판별형 기반 레이다-카메라 융합 모델들과의 비교를 통해 Stable Diffusion의 사전 지식의 영향도 함께 확인하고자 한다.

3-2 주요 결과

표 2는 기존 단일 카메라 기반 생성형 베이스라인 모델 및 대표적인 레이다-카메라 융합 판별형 모델들과 본 논문이 제안한 최종 모델의 성능을 정량적으로 비교한 결과이다. 제안된 모델은 베이스라인 모델 대비 0∼80 m에서의 MAE와 RMSE가 각각 30.5 %, 12.6 % 감소하는 모습을 보인다. 판별형 모델들과의 정량적 비교에서는 레이다 포인트의 고밀도 보간에 치중된 판별형 모델들이 우세한 지표를 나타낸다.

그러나 이들 기법은 레이다 포인트가 도달하지 않는 픽셀 영역에서 추정 경계선이 무너지거나 깊이 맵의 공간 해상도가 크게 저하되어 시각적으로 거친 아티팩트를 양산하는 고질적인 한계를 안고 있다. 현재의 표준 평가 프로토콜은 희소한 라이다 포인트가 존재하는 픽셀에서만 오차를 계산한다. 즉 MAE와 RMSE에는 전체 이미지 중 라이다 감독 신호가 부재한 대다수의 픽셀 영역에서의 추정 품질은 평가에 전혀 반영되지 않는다. 이러한 문제는

표 2. 깊이 추정 성능 결과 정량적 비교
Table 2. Quantitative comparison of each model.

Distance	Model	MAE (m)	RMSE (m)
0∼50 m	Marigold[2]	4.95	7.37
	Ours	2.63	4.58
	Singh et al.[4]	1.73	3.75
	CaFNet[5]	1.67	3.67
	TacoDepth[7]	1.42	3.25
0∼70 m	Marigold[2]	5.27	7.90
	Ours	3.46	6.38
	Singh et al.[4]	2.07	4.59
	CaFNet[5]	2.01	4.49
	TacoDepth[7]	1.71	3.96
0∼80 m	Marigold[2]	5.35	8.07
	Ours	3.72	7.05
	Singh et al.[4]	2.18	4.90
	CaFNet[5]	2.11	4.77
	TacoDepth[7]	1.83	4.15

3D 객체 검출 또는 경로 계획과 같은 자율주행 하위 작업에 활용되었을 때 예측 정확도가 떨어지거나 주행 안정성에 위험을 초래할 수 있다[10]. 실제로 그 문제는 그림 2에서 나타난다. 판별형 모델들은 깊이 추정 경계선이나 구조적 정보가 무너져 시각적으로 거친 아티팩트를 보인다. 반면, 제안된 모델은 대규모 데이터셋으로 사전 학습된 Stable Diffusion의 잠재 기하학적 사전 지식을 융합 네트워크의 기반으로 삼는다. 이를 통해 그림 2의 정성적 시각 결과에서 타 판별형 모델들 대비 상대적으로 고화질의 깊이 추정 성능을 보여준다. 특히 객체의 경계 영역이 보다 선명하고, 라이다 ground truth가 정답을 제공하지 못하는 부분의 깊이까지 복원해 낸다.

$$F-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

$$Spearman = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \tag{6}$$

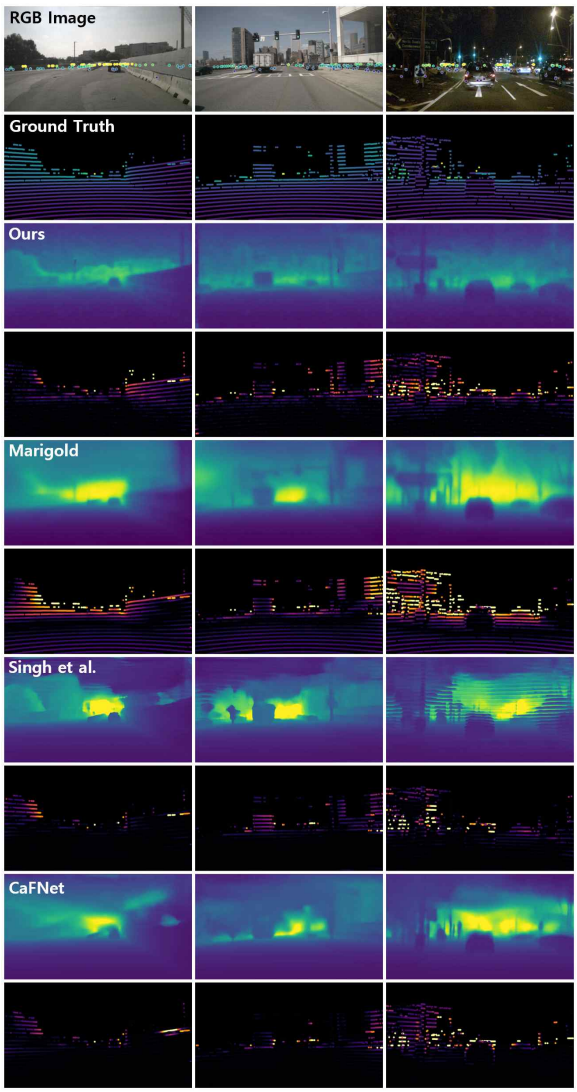


그림 2. 깊이 추정 성능 결과 정성적 비교
Fig. 2. Qualitative comparison of each model.

또한 표 3에서는 이를 정량적 지표로 보여준다. 식 (5)의 precision은 각 모델 결과의 경계 중 Marigold 결과의 경계와 일치하는 비율, recall은 Marigold 결과의 경계 중 각 모델 결과의 경계와 일치하는 비율이다. 또 식 (6)의 R(X)와 R(Y)는 각각 Marigold와 각 모델 결과의 픽셀별 깊이 값 순위를 나타낸다. 확산 모델을 기반으로 고화질의 깊이 맵을 생성하는 최신 모델인 Marigold를 기반으로 생성한 깊이 맵을 기준으로 경계선에 대한 F-score 및 픽셀 별 깊이 값의 순서에 대한 Spearman 상관계수를 계산

표 3. 모델별 경계와 구조 정보 복원 성능
Table 3. Model-specific edge and structural information restoration performance.

Model	F-score ↑	Spearman ↑
Singh et al.[4]	0.415	0.856
CaFNet[5]	0.534	0.892
Ours	0.620	0.921

한 결과, 제안된 모델이 기존 판별형 모델보다 더 좋은 경계 복원 성능과 구조 정보 복원 성능을 가지는 것을 알 수 있다. 이는 제안된 모델이 레이다의 깊이 정보를 확산 모델의 사전 지식과 상호보완적으로 정렬함으로써, 단일 카메라 깊이 추정 모델의 절대 깊이 스케일 모호성을 해결하는 동시에 시각적 완성도가 향상된 깊이 맵을 출력할 수 있음을 제시한다.

제안 모델과 기존 판별형 모델들의 정량적 격차는 향후 개선되어야 할 과제이다. 다만 현재의 희소 라이다 포인트 기반 평가 지표가 깊이 맵 품질의 전부를 대변하지 못한다는 점, 그리고 그 지표가 포착하지 못하는 구조적 완성도가 자율주행 시스템의 안전성과 직결된다는 점을 고려할 때, 본 접근은 기존 판별형 패러다임을 보완하는 유의미한 방향임을 보여준다. 아울러 경계부 정확도나 구조적 품질을 정량화할 수 있는 평가 지표를 도입하여 이러한 우위를 수치로 입증하는 것 또한 흥미로운 후속 연구가 될 것이다.

3-3 Stable Diffusion 사전 지식 절제 실험 결과

Stable Diffusion은 대규모 데이터셋으로 사전 학습되어 현실 세계의 원근감, 객체의 물리적 경계 등에 대한 지식을 이미 보유하고 있다. 본 제안 모델이 Stable Diffusion의 사전 지식을 실질적으로 수혜받고 있는지 검증하기 위해, 전체 Stable Diffusion 백본과 동일 아키텍처(full) 및 제안 모델과 학습 파라미터 크기를 맞춘(matched) 조건에서 무작위 초기화 가중치로 처음부터 학습한 대조군 실험을 진행하였고, 표 4와 그림 3은 그 결과를 보여준다.

동일한 U-Net 아키텍처를 적용했음에도 무작위 초기화 가중치로 처음부터 학습하는 경우 성능이 악화되는 양상을 나타낸다. 무작위 초기화 가중치로 처음부터 학습을

표 4. nuScenes 0~80 m에 대한 Stable Diffusion 사전 지식 절제 실험 정량적 결과
Table 4. Stable diffusion prior ablation quantitative results on 0~80 m of nuScenes dataset.

Type	MAE (m)	RMSE (m)
Scratch (full)	7.95	11.58
Scratch (matched)	7.85	12.63
Ours	3.72	7.05

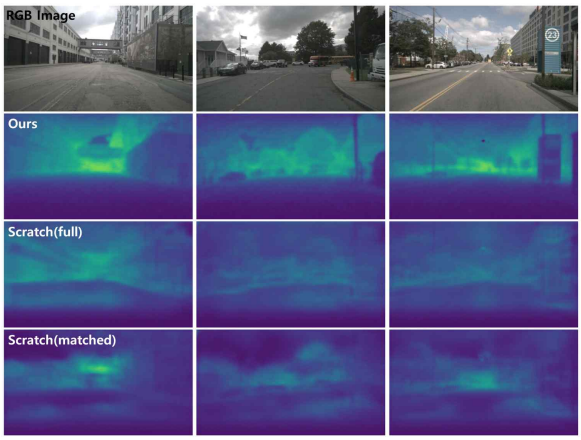


그림 3. nuScenes 0~80 m에 대한 Stable Diffusion 사전 지식 절제 실험 정성적 결과
Fig. 3. Stable Diffusion prior ablation qualitative results on 0~80 m of nuScenes dataset.

시도하는 경우, 모델이 이미지의 기본 특징부터 3D 공간적 정렬 관계까지 동시에 학습해야 한다. 하지만 한정된 자율주행 데이터셋만으로는 이를 정교하게 수렴시키기 힘들고, 결국 성능 악화로 이어지게 된다.

3-4 제안 기법 절제 실험 결과

표 5는 본 논문에서 제안한 세 가지 핵심 아키텍처 설계 선택지의 타당성을 검증하기 위한 절제 실험 결과를 보여준다. 확산 모델 특유의 다단계 노이즈 스케줄을 제거하고 판별 모델 기반의 단일 패스 회귀 구조로 전환했을 때 성능이 유의미하게 향상된다. 이는 매 단계 추가되는 가우시안 노이즈가 레이다가 제공하는 확실한 절대 깊이 신호를 희석하고 방해했음을 시사하며, 픽셀 공간에

표 5. nuScenes 0~80 m에 대한 절제 실험 결과
Table 5. Ablation results on 0~80 m of nuScenes dataset.

Single-pass prediction	Initialized attention	Frozen U-Net	MAE (m)	RMSE (m)
×	×	×	4.78	8.86
✓	×	×	4.32	8.13
✓	✓	×	4.11	7.57
✓	✓	✓	3.72	7.05

서의 손실 함수를 통한 학습이 효과적임을 증명한다.

또한 기존 Stable Diffusion의 텍스트 어텐션 가중치를 사용하지 않고 무작위 초기화된 고유의 레이다 전용 어텐션을 도입하면서 MAE가 4.32 m에서 4.11 m로 감소하는 모습을 보인다. 그림 4는 U-Net의 마지막 레이어에서 기존의 텍스트 어텐션 가중치를 가진 모델과 무작위 초기화된 어텐션 모델에서 특정 레이다 포인트와 주변 이미지 픽셀과의 어텐션 점수를 시각화한 결과이다. 그림 4에서 알 수 있듯이, 무작위 초기화된 레이다 전용 어텐션 모델이 레이다 포인트와 관련 있는 이미지 픽셀에 더 강하게 집중하는 모습을 보여준다. 언어 도메인의 표현 공간에 결속되어 있던 기존 가중치의 영향을 제거함으로써, 통계적 특성이 다른 레이다 특징 벡터가 이미지 잠재 공간과 정렬될 때 발생하는 오류를 차단했기 때문이다.

마지막으로 U-Net 백본을 완전히 동결하고, 레이다에

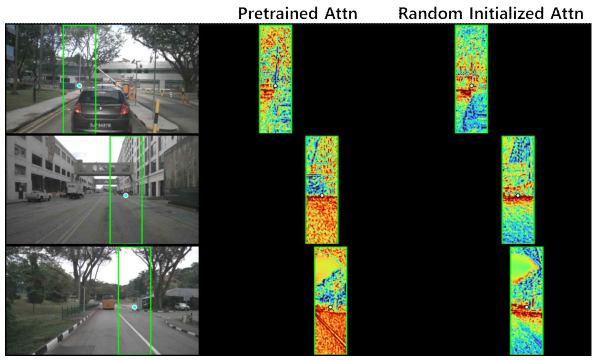


그림 4. 기존 및 신규 크로스 어텐션 모델의 어텐션 맵 시각화
Fig. 4. Visualization of attention map of pretrained and randomly initialized cross-attention model.

표 6. 레이다 노이즈 대체 시 U-Net 백본 동결에 따른 성능 하락 폭 비교 (0~80 m MAE)

Table 6. Comparison of performance drop due to U-Net backbone freezing when replacing radar with noise.

Type	Radar depth	Noise depth	ΔMAE
Trainable U-Net	4.11	4.25	+0.14
Frozen U-Net	3.72	3.96	+0.24

대한 좁은 하위 네트워크만 학습시켰을 때 MAE가 4.11 m에서 3.72 m의 성능 향상을 보인다. 특히 표 6에서 볼 수 있듯이, 레이다 포인트의 깊이 정보를 노이즈로 바꿔 넣어주었을 때 U-Net 백본을 동결한 모델이 그렇지 않은 모델보다 성능 하락 폭이 더 크다. 이는 U-Net 백본을 동결한 모델이 깊이를 추정하는 데 있어 레이다의 정보를 보다 적극적으로 활용한다는 지표이다. 백본의 거대한 용량을 열어둘 경우 모델이 레이다 경로를 무시하고 이미지에서 곧장 깊이로 가는 지름길을 학습하는 현상이 발생하지만, 백본 코어를 동결함으로써 네트워크가 정확한 깊이 추정을 위해 레이다 신호를 적극적으로 활용하게 되었음을 보여준다.

Ⅳ. 결 론

본 논문에서는 대규모 이미지 데이터로 사전 학습된 생성형 확산 모델을 레이다-카메라 융합 깊이 추정 작업에 성공적으로 이식하기 위한 세 가지 핵심 설계 선택지를 제안하고, 그 유효성을 검증한다. 다단계 노이즈 스케줄 대신 단일 패스 추정을 도입하고, 텍스트 어텐션 가중치를 제거하여 레이다에 맞게 학습할 수 있도록 한다. 또 좁은 하위 네트워크만 학습하여 이미지의 영향을 너무 과하게 받지 않도록 하고 레이다 신호에 대한 학습을 유도한다.

본 제안 기법은 기존 판별형 융합 모델들에 비해 정량적 성능이 낮아 앞으로 더 개선해야 할 필요가 있다. 하지만 본 제안 기법은 사전 학습된 생성 모델의 풍부한 잠재 기하학적 사전 지식을 활용함으로써 상대적으로 고화질의 깊이 맵을 추정해 낸다. 선명한 경계를 갖고 노이즈가 제거된 깊이 정보는 자율주행 하위 작업 전체의 주행 안정성과 시스템 신뢰성을 확보하는 데 중요한 역할을 할 것이다.

본 연구는 대규모 사전 학습된 생성 모델을 완전히 새로운 모달리티가 개입하는 조건부 하위 작업으로 확장할 때, 기존 가중치를 재사용하고 전체 네트워크를 파인튜닝하며 원래의 확산 패러다임을 그대로 유지하는 것이 항상 최선의 선택은 아니라는 점을 보인다. 오히려 대상 작업의 물리적 특성에 부합하는 직접적인 손실 함수를 정의하고, 모달리티에 맞는 설계를 도입하는 것이 좋은 성능과 안정성을 얻을 수 있다. 또한 본 논문에서는 레이다의 3차원 공간 좌표 정보만을 사용하였으나, 향후 연구에서는 레이다 센서 고유의 풍부한 물리적 특성인 도플러 신호나 RCS 정보를 추가로 활용함으로써 객체의 특성과 레이다의 신뢰도를 깊이 추정에 반영하고 동적 객체의 감지 성능을 고도화할 여지가 충분히 존재한다. 아울러 지식 증류를 활용한 모델 경량화의 방향으로도 확장이 가능할 것으로 판단된다.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, New Orleans, LA, Jun. 2022, pp. 10674-10685.

[2] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Seattle, WA, Jun. 2024, pp. 9492-9502.

[3] J. S. Kim, W. H. Park, P. H. Yoon, B. S. Kim, and R. Song, "Object detection using stereo vision and radar sensor fusion," *The Journal of Korean Institute of Electromagnetic Engineering and Science*, vol. 31, no. 11, pp. 986-989, Nov. 2020.

[4] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, and S. Soatto, et al., "Depth estimation from camera image and mmWave radar point cloud," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), Vancouver, BC, Jun. 2023, pp. 9275-9285.

[5] H. Sun, H. Feng, J. Ott, L. Servadei, and R. Wille, "CaFNet: A confidence-driven framework for radar camera depth estimation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS)*, Abu Dhabi, Oct. 2024, pp. 2734-2740.

[6] H. Li, Y. Ma, Y. Gu, K. Hu, Y. Liu, and X. Zuo, "RadarCam-Depth: Radar-camera fusion for depth estimation with learned metric scale," in *2024 IEEE International Conference on Robotics and Automation(ICRA)*, Yokohama, May 2024, pp. 10665-10672.

[7] Y. Wang, J. Li, C. Hong, R. Li, L. Sun, and X. Song, et al., "TacoDepth: Towards efficient radar-camera depth estimation with one-stage fusion," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, Jun. 2025, pp. 10523-10533.

[8] A. Wong, X. Fei, S. Tsuei, and S. Soatto, "Unsupervised depth completion from visual inertial odometry," in *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1899-1906, Apr. 2020.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, and Q. Xu, et al., "nuScenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, Seattle, WA, Jun. 2020, pp. 11618-11628.

[10] H. Li, M. Jing, W. Jin, S. Dong, J. Liang, H. Fan, and R. Ji, "Sparse beats dense: Rethinking supervision in radar-camera depth completion," in *Computer Vision – ECCV 2024: 18th European Conference, Proceedings, Part XLIX*, Milan, Sep.-Oct. 2024.

이 재 룡 [대구경북과학기술원/학사과정]

https://orcid.org/0009-0005-2869-5910



2020년 2월~현재: 대구경북과학기술원 기초학부 학사과정

[주 관심분야] Multi-Sensor Fusion, 컴퓨터 비전 등

신 정 완 [대구경북과학기술원/박사후연구원]

https://orcid.org/0000-0002-6990-2892



2015년 2월: 영남대학교 정보통신공학과 (공학사)

2018년 2월: 경북대학교 컴퓨터학부 (공학석사)

2025년 8월: 경북대학교 컴퓨터학부 (공학박사)

2025년 8월~현재: 대구경북과학기술원 전기전자컴퓨터공학과 박사후연구원

[주 관심분야] 자연어처리, 기계학습 등

신 헌 용 [국방과학연구소/선임연구원]

https://orcid.org/0009-0006-7247-4215



2007년 2월: 부산대학교 전자전기통신공학과 (공학사)

2009년: 광주과학기술원 정보통신공학과 (공학석사)

2012년~2021년: 국방기술품질원 선임연구원

2021년~현재: 국방과학연구소 선임연구원

[주 관심분야] 전투체계, 심층 강화학습, 무인수상정 경로계획

김 동 영 [국방과학연구소/선임연구원]

https://orcid.org/0009-0003-4912-299X



2011년: Iowa State University Computer Science (B.S.)

2016년: Iowa State University Human-Computer Interaction (M.S.)

2013년~2017년: 대구경북과학기술원 IT융합연구부 연구원

2017년~현재: 국방과학연구소 선임연구원

[주 관심분야] 전투체계, 심층 강화학습, 무인수상정 경로계획

김 민 [국립부경대학교/조교수]

https://orcid.org/0000-0001-7520-3513



2015년 2월: 국립부경대학교 전자공학과 (공학사)

2017년 2월: 국립부경대학교 전자공학과 (공학석사)

2022년 8월: 포항공과대학교 전자전기공학과 (공학박사)

2023년 3월～2026년 9월: 한국해양과학기술원 해양 ICT · 모빌리티 연구부 선임연구원

2026년 9월～현재: 국립부경대학교 스마트모빌리티공학과 조교수

[주 관심분야] 레이다 신호처리, 레이다 추적, 레이다 표적 식별, 비접촉식 보행/생체신호 센싱

최 재 호 [대구경북과학기술원/조교수]

https://orcid.org/0000-0001-9484-4869



2017년 2월: 고려대학교 컴퓨터·통신공학부 (공학사)

2019년 2월: 포항공과대학교 전자전기공학과 (공학석사)

2023년 2월: 포항공과대학교 전자전기공학과 (공학박사)

2023년 4월～2024년 12월: 미국 Stanford University 전기전자공학과 박사후연구원

2025년 3월～현재: 대구경북과학기술원 전기전자컴퓨터공학과 조교수

[주 관심분야] 레이다 AI, 레이다 신호처리, 무선 센싱, 센서 융합 등