2024

FINAL REPORT

단안 깊이 추정으로 3D 재구성

22013574 인공지능학과 최지훈

목차

01. 개요

깊이 추정 프로젝트의 목적

깊이 추정 프로젝트는 2D 이미지나 비디오에서 3D 공간 정보를 획득하여 다양한 분야에 활용됩니다. 이를 통해 자율주행, 로봇 제어, AR/VR, 의료 영상 처리, 영화 및 게임 제작, 지리 공간 분석, 소비자 애플리케이션 등에서 3D 환경 재구성과 거리 계산을 가능하게 합니다.

깊이 추정 방법

깊이 추정은 2D 데이터에서 3D 공간 정보를 복원하기 위한 기술로, 주요 방법은 단안, 멀티 뷰, LiDAR 기반 방법들이 있습니다.

1. 단안 기반 방법 (Monocular Depth Estimation)
단안 방법은 단일 이미지로부터 깊이를 추정하는 기술로, 추가적인 장치 없이 활용 가능하다. 기하학적 단서(선형 원근법, 물체 크기 등)나 학습 기반 모델을 사용해 깊이를 계산한다.

2. 멀티 뷰 기반 방법 (Multi-View Depth Estimation)
멀티 뷰 방식은 여러 각도에서 촬영한 이미지를 사용하여 깊이를 계산한다. 스테레오 비전(양안 시차), 다중 뷰 기하학(SfM) 등을 활용해 높은 정확도를 제공한다.

3. LiDAR 기반 방법
LiDAR(Light Detection and Ranging)는 레이저를 사용해 물체까지의 거리를 측정하는 방식으로, 깊이 정보를 직접 획득한다. 이는 실시간 3D 맵핑 및 고정밀 거리 측정에 주로 사용된다.

단안 카메라 선택 이유

깊이 추정 방법 중 단안 카메라를 선택한 이유는 멀티 뷰와 LiDAR 기반 방법이 각각 카메라의 개수 및 배열, 센서의 정밀도에 크게 의존하며, LiDAR의 경우 빛 반사의 특성에 따라 오류가 발생할 가능성이 있기 때문입니다. 반면, 단안 카메라는 상대적으로 저렴한 비용으로 사용할 수 있으며, 다른 센서와 결합하여 활용할 경우 유연성과 확장성이 더 뛰어나다고 판단됩니다.

02. 데이터셋

NYU DEPTH V2

NYU Depth V2 데이터셋은 실내 장면에서 RGB-D 데이터를 제공하는 가장 널리 사용되는 데이터셋 중 하나로, 뉴욕대학교(New York University)의 연구진이 생성한 데이터셋입니다. 이 데이터셋은 컴퓨터 비전, 특히 심도 예측(depth prediction), 실내 장면 분할(scene segmentation), 물체 인식(object recognition), 및 3D 복원(3D reconstruction)과 같은 응용 분야에서 중요한 역할을 합니다.

이 데이터셋을 선택한 이유는 단안 카메라가 외부 환경 학습 시 복잡성과 거리 검출의 한계가 있을 것으로 예상되었기 때문입니다. 따라서 내부 데이터셋을 활용하여 단안 카메라의 한계를 극복하고 더 발전된 깊이 추정 능력 발전 시킬수 있다고 생각했습니다.

그림 2.1 NYU DEPTH V2 데이터셋 형태

NYU DEPTH V2 데이터 형태

NYU Depth V2 데이터셋은
RGB이미지와 대응하는 깊이 데이터(Depth map)를 포함
실내 환경에서 수집된 약 1,449개의 고해상도 RGB-D 이미지로 구성
추가적으로, 약 407,024개의 저해상도 프레임이 비디오 데이터 형태

RGB 이미지: 640 x 480 해상도의 컬러 이미지.
Depth Map: Microsoft Kinect 센서를 사용해 수집된 640 x 480 해상도의 심도 데이터.

02. 데이터셋

데이터셋 전처리

비전 이미지 처리에서 가장 많이 사용되는 필터는 가우시안 필터와 칼만 필터입니다. 이번 프로젝트에서는 특징을 더 잘 추출하기 위해, 서로 다른 시그마 값을 적용한 가우시안 필터를 빼는 DoG(Difference of Gaussians) 기법을 활용하였으며, 이를 변형한 저만의 DoG 방식을 개발하여 사용하였습니다.

아래 이미지[그림 2.2]의 Focus Points를 보면 6×6 그리드에서 가장자리 부분을 제외한 4×4 그리드의 점들이 표시되어 있습니다. 총 16개의 점 중 각 점 주변의 분산을 계산하여, 분산 값이 높은 점을 탐지하였습니다. 분산 값이 높은 점을 찾는 이유는 초점과 비슷한 점을 식별하여 흐림 효과를 적용하기 위함입니다. 이를 가우시안 DoG 기법과 결합하기 위해 가장 큰 분산값과 그다음으로 큰 분산값의 차이를 시그마 값으로 설정하여 필터를 구현하였습니다.
이 방식은 기존의 DoG 기법을 확장하여, 초점 영역을 더욱 정밀하게 처리하는 데 초점을 맞추었습니다. [그림 2.3]을 보면 평균적으로 1장 처리하는 속도는 0.02초입니다.
* 추가적으로 분산값들의 차이가 0이거나 1보다 작은 차이를 가질 경우 2로 고정하여 오류가 뜨는 경우를 삭제했습니다.

그림 2.2 변형 Difference of Gaussians 방법

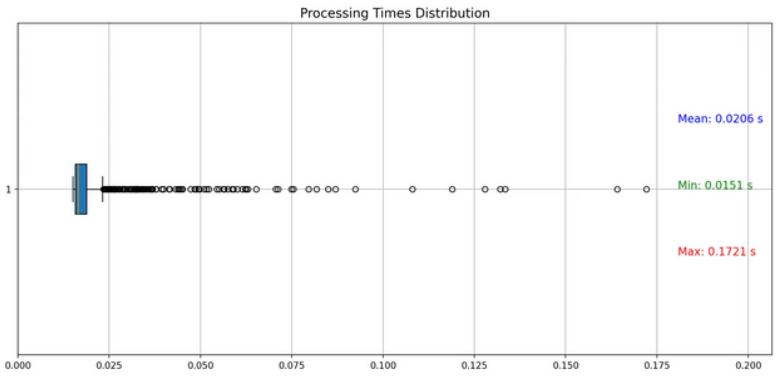

그림 2.3 변형 Difference of Gaussians 처리 시간

02. 데이터셋

입력 데이터로 전처리

이전에 제작한 개량 DoG 필터와 원본 이미지를 모델 이미지에 넣기 전에 합치는 과정을 추가했습니다. 합성 과정은 [그림 2.4]에 나타나 있습니다.

합성 과정에서는 480 x 640 x 3 크기의 원본 이미지와 480 x 640 x 1 크기의 블러 이미지를 사용합니다. 블러 이미지를 원본 이미지와 결합하기 위해, 먼저 블러 이미지의 차원을 480 x 640 x 3으로 확장합니다.
그런 다음, 이 두 이미지를 파이토치의 concat 과정을 통해 결합하여 480 x 640 x 6의 최종 형태로 만들어 학습 모델에 입력할 데이터를 생성합니다.
이러한 과정을 통해 원본 이미지보다 더 나은, 나뉜 영역별 깊이 정보를 얻을 수 있다고 판단했습니다. 저는 단안 카메라로는 세세한 깊이 정보를 얻기보다는 영역별로 명확한 깊이 데이터를 추출하는 것이 적합하다고 생각했습니다. 이는 현재의 트렌드가 세부 깊이 정보는 고성능 깊이 측정 센서를 사용하는 방향으로 이동하고 있기 때문입니다.

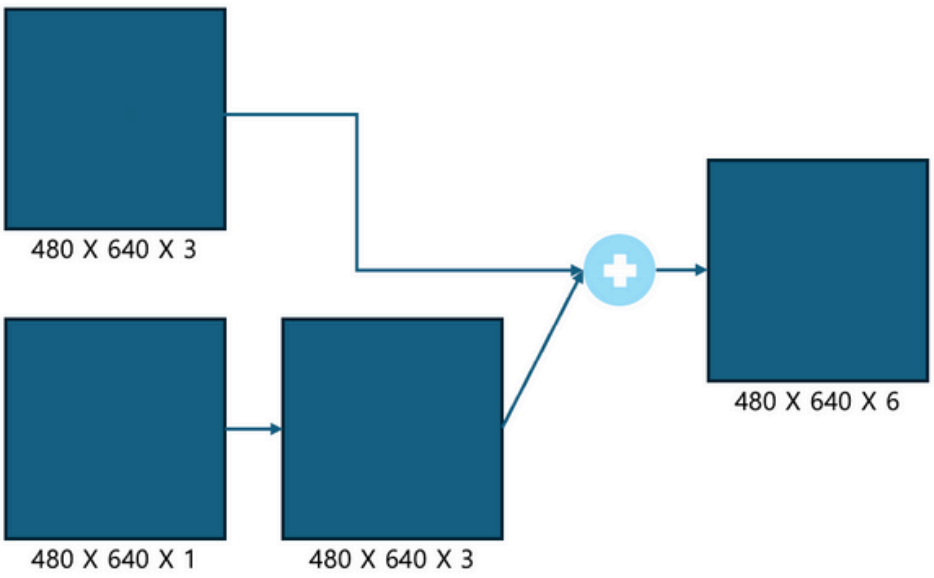

그림 2.4 원본이미지와 블러 이미지 합성방법

03. 모델

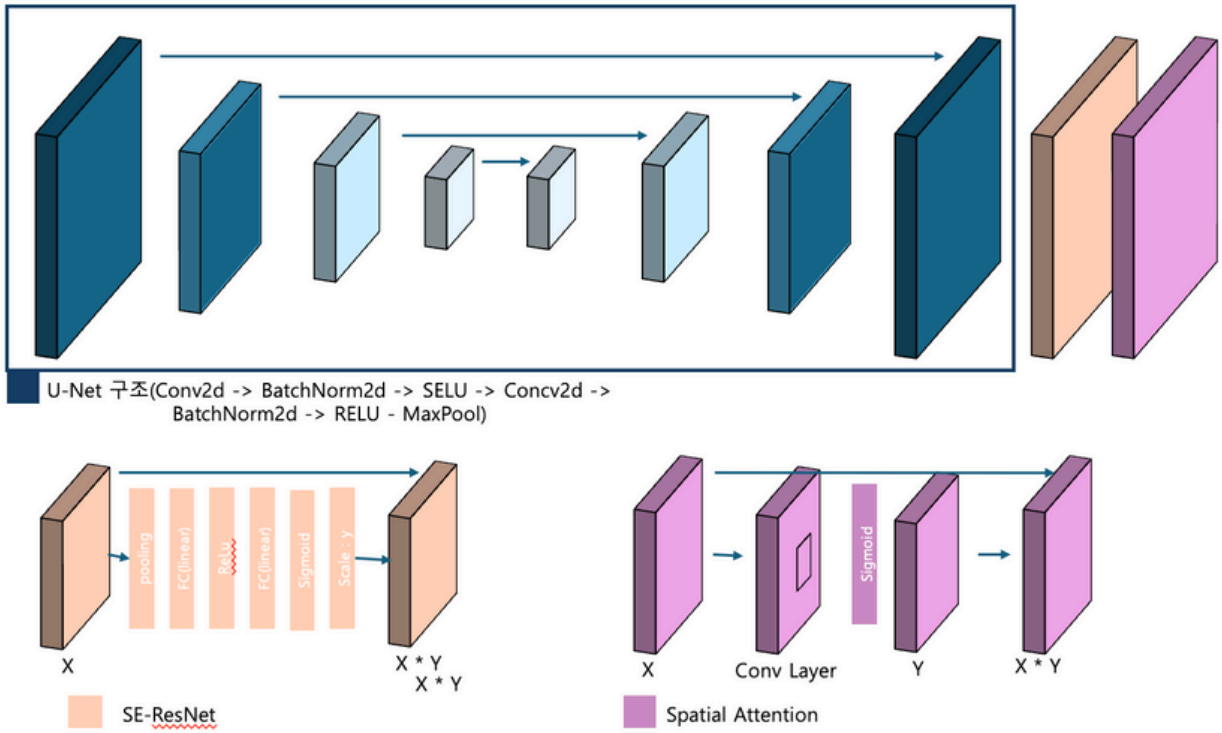

그림 3.1 전체적인 모델 구조

모델 구조

단안 깊이 추정에서 가장 널리 사용되는 모델 구조의 트렌드는 U-Net, 어텐션, 그리고 트랜스포머라고 판단했습니다. 이를 바탕으로, 저는 기본적인 U-Net 구조를 따르되, 활성화 함수는 RELU에서 SELU로 변경하였습니다.
일반적으로 U-Net 뒤에 Vision Transformer(ViT)나 어텐션 모델을 추가하여 더 강건한 모델을 만드는 경우가 많습니다. 그러나 저는 자원 소모를 줄이면서도 성능을 향상시키기 위해 트랜스포머 대신 어텐션 층을 두 개 포함하는 모델을 설계했습니다. 이를 통해 더 가벼우면서도 효율적인 모델을 구현하고자 했습니다.
추가적으로, 모델을 더 가볍게 만들기 위해 트랜스포머를 사용하지 않았습니다. 대신, 이 단안 깊이 추정 모델이 다른 센서들과의 협업을 염두에 두고 설계되었기 때문에, 대략적인 영역의 깊이를 정확하게 측정하는 데 중점을 두었습니다.
따라서 모델 구조는 특정 환경에 과적합되지 않고, 다양한 상황에서 일반화 성능을 잘 발휘할 수 있는 방향으로 설계하였습니다. 이러한 접근을 통해 효율성과 활용성을 동시에 고려한 모델을 구현하고자 했습니다.

전체적인 제가 이번 학습에 사용한 모델 구조는 [그림3.1]과 같습니다.

03. 모델

최적화 방법 및 로스함수

이번 학습 모델을 설계하면서 가장 중점을 둔 부분은 손실 함수(Loss Function)의 정의였습니다. 일반적으로 깊이 추정 모델에서 많이 사용되는 손실 함수는 L1 Loss로, 추정된 깊이 맵과 실제 깊이 맵 간의 절대적인 차이를 계산하는 방식입니다. L1 Loss는 깊이 추정의 정확도를 보장하기 위한 기본적인 손실 함수로 활용됩니다. 하지만 단순히 픽셀 단위 차이만 고려하는 L1 Loss의 한계를 보완하고자, SSIM(Structural Similarity Index Measure)을 추가로 사용하였습니다. SSIM은 추정된 깊이 맵과 실제 깊이 맵 간의 구조적 유사성을 비교하여 깊이 맵의 시각적 품질과 구조적인 세부 사항을 보존하는 데 기여합니다. 이를 통해 L1 Loss가 놓칠 수 있는 세부적인 구조를 보완할 수 있었습니다.
또한, 깊이 맵에서 깊이가 증가하거나 감소하는 경향성을 학습하기 위해 Gradient Loss(경사 손실)를 추가하였습니다. Gradient Loss는 깊이 변화가 급격히 발생하는 영역에서 경계와 패턴을 효과적으로 학습하도록 도와줍니다. Gradient Loss를 설계하는 과정에서는 전역 경사 손실(Global Gradient Loss)과 그리드 경사 손실(Grid-based Gradient Loss) 두 가지 접근 방식을 사용했습니다. 전역 경사 손실은 깊이 맵 전체의 경사 정보를 계산하여 전반적인 패턴을 학습하는 데 유용합니다. 반면, 그리드 경사 손실은 깊이 맵을 4x4 크기의 그리드로 나누고, 각 그리드에서 경사 손실을 계산한 후 이를 L1 Loss로 합산하여 국소적인 세부 정보를 학습하는 방식입니다.
두 가지 경사 손실을 결합하는 과정에서, 그리드 경사 손실이 과도하게 반영되면 모델의 일반화 성능이 저하될 수 있다는 점을 고려하여 그리드 경사 손실의 가중치를 0.3으로 설정하였습니다. 이는 전역 경사 손실과 그리드 경사 손실 간의 균형을 맞추기 위한 조치로, 전역적인 패턴과 국소적인 변화 모두를 효과적으로 학습하도록 설계한 것입니다

최종 손실 함수는 L1Depth + Grad_global + 0.3 * Grad_grid + SSIM으로 설계하였습니다. 이 손실 함수는 깊이 추정의 정확성과 시각적 품질을 모두 고려할 수 있도록 구성되었습니다. 옵티마이저로는 Adam을 사용하였으며, 학습률은 0.009로 설정하였습니다.
학습에 사용된 데이터는 전체 5만 개 중 2만 개를 선택하였으며, 배치 사이즈는 30으로 설정하였습니다. 학습 에포치는 50으로 설정하였고, 50번째 에포크에서 train loss = 4.9, val_loss = 5.7으로 가장 성능이 좋았습니다.
학습은 Tesla V100 (32GB VRAM)을 사용하여 진행되었으며, 학습 중 VRAM 사용량은 31GB였습니다. 총 학습 시간은 약 28시간이 소요되었습니다.

04. 모델 성능 분석

성능분석

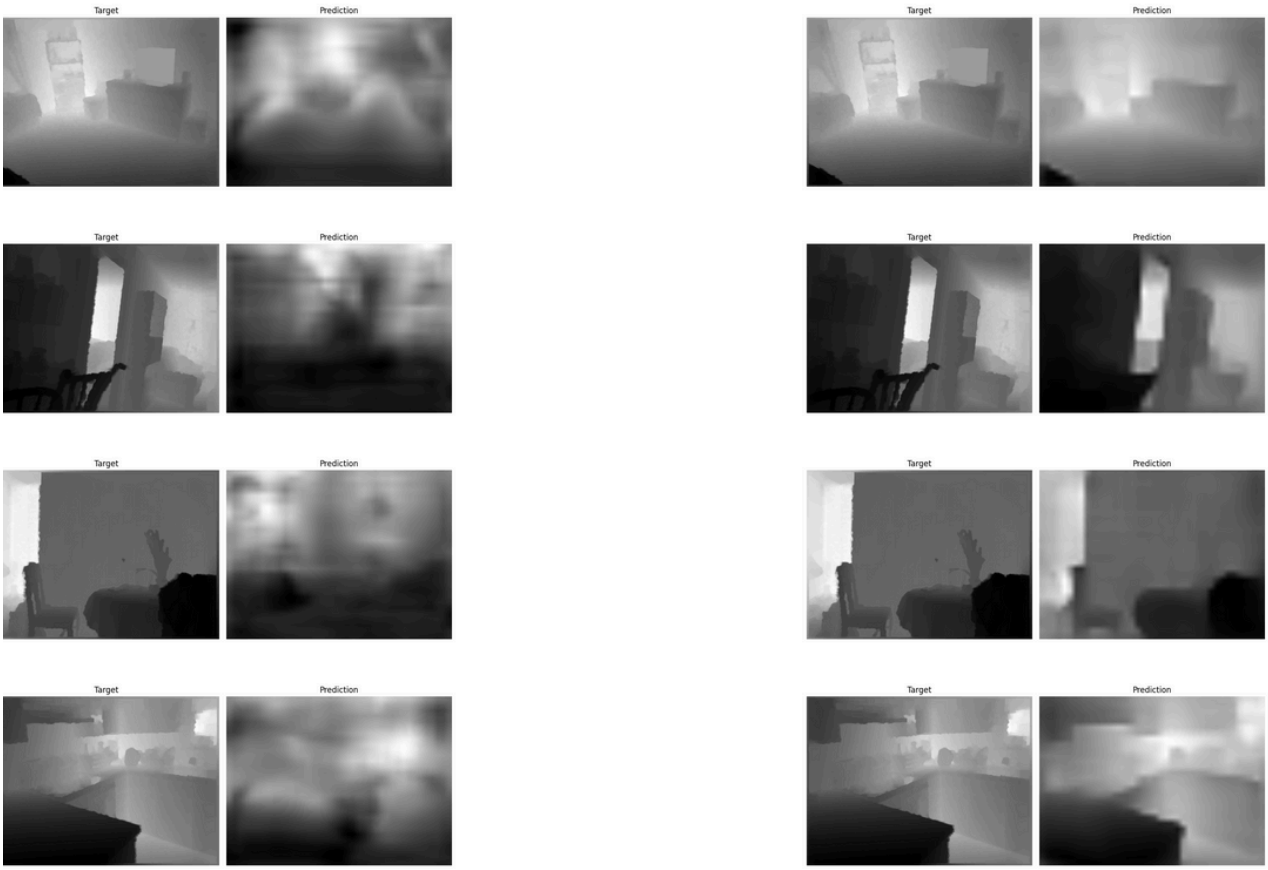

그림 4.1 에포치 1 -> 에포치 2 벨리데이션 셋 변화

위 [그림 4.1]은 에포치 1과 에포치 50에서의 벨리데이션 셋을 시각화한 결과입니다. 에포치 1일 때 로스 함수 값은 72.4였으며, 에포치 50에서는 점차 감소하여 5.7로 가장 낮은 값을 기록했습니다. 시각화 결과를 보면, 모델이 대체로 영역별로 깊이 정보를 잘 반영한 것을 확인할 수 있습니다.

랜덤한 100장을 뽑아서 성능 분석
Average Losses:
  depth_loss: 3.3032
  grad_global_loss: 0.7557
  grad_grid_loss: 0.6236
  ssim_loss: 0.9876
  total_loss: 5.2337

05. 모델 배포

모델 배포 FLASK

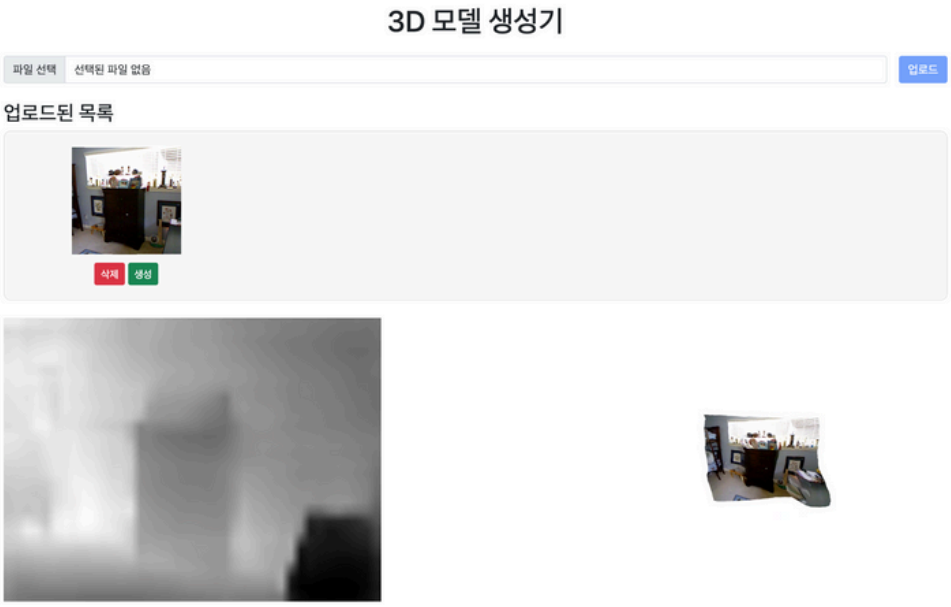

그림 5.1 웹사이트 화면

Flask를 사용하여 웹사이트에서 모델 API를 간편하게 구현했습니다. 업로드한 이미지 아래에 있는 '생성' 버튼을 클릭하면, 깊이 맵과 함께 360도 회전하는 영상이 생성됩니다. 깊이 맵 생성에는 약 1.3초가 소요되는 반면, 영상 생성은 약 1분 30초 정도 걸립니다(MacBook M1 기준).

github : https://github.com/choimagon/monodepthFinal

06. 참고문헌

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv. https://arxiv.org/abs/1505.04597

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2018). Squeeze-and-excitation networks. arXiv. https://arxiv.org/abs/1709.01507

Chen, C., Gong, D., Wang, H., Li, Z., & Wong, K. K. (2020). Learning spatial attention for face super-resolution. arXiv. https://arxiv.org/abs/2012.01211

Jung, G., & Yoon, S. M. (n.d.). Single image-based depth estimation network using attention model. HCI Lab., Computer Science, Kookmin University.