전기전자종합설계 및 소프트웨어 실습 최종보고서

<table>
<tr><td>연구
주제</td><td colspan="5">단안 카메라 깊이 추정 모델 (Lite-mono) 성능 개선</td></tr>
<tr><td>연구
팀명</td><td colspan="5">한승민(개인)</td></tr>
<tr><td>연구
기간</td><td colspan="5">2025년 01월 20일 ~ 2025년 06월 15일</td></tr>
<tr><td rowspan="2">지도
교수</td><td>소속</td><td colspan="2">전기전자공학부</td><td>제출일자</td><td>2025. 11. 12</td></tr>
<tr><td>성명</td><td colspan="2">김원준</td><td>확인란</td><td>서명</td></tr>
<tr><td rowspan="2">팀장</td><td>소속</td><td>학번</td><td>이름</td><td>연락처</td><td>이메일</td></tr>
<tr><td>한승민</td><td>202011068</td><td>한승민</td><td>01071284987</td><td>seungmin4987
@naver.com</td></tr>
<tr><td rowspan="5">팀원</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
</table>

특이 사항	

목 차

0. 요약문

본 연구에서는 경량 단안 깊이 추정 모델인 Lite-Mono를 기반으로, 자기지도방식의 단안 깊이 추정(Self-Supervised Monocular Depth Estimation) 모델의 성능을 향상시키기 위한 여러 방안을 제안한다.

먼저 KITTI 데이터셋의 eigen_zhou split을 이용해 Lite-Mono 기본 모델을 재현하고, 이미지 재구성 손실과 스무딩 손실로 구성된 멀티 스케일 손실 함수를 적용하여 기준 성능을 확보하였다. 이후 성능 향상을 위해 (1) 엔코더에 Spatial Attention 모듈을 추가, (2) Depth Anything V2를 Teacher 네트워크로 활용한 pseudo label 기반 하이브리드 학습, (3) AdaBins 아이디어를 응용한 disparity binning + Chamfer Distance Loss 기법을 적용하여 비교하였다.

그 결과, 제안한 disparity binning 기반 모델은 기준 Lite-Mono 대비 오차 지표(abs_rel, sq_rel, rmse, rmse_log)를 전반적으로 개선하면서, 경량성을 크게 해치지 않는 수준의 FLOPs 증가만을 보였다.

1. 서론

1.1 본 연구의 동기 및 필요성

▲그림1. 단안카메라 기반 깊이 추정 예시

단안 깊이 추정은 카메라 한 대만으로 장면의 3차원 구조를 유추하는 기술로, 자율주행·로봇 내비게이션·증강현실 등 다양한 응용 분야에서 핵심적인 역할을 한다.

최근에는 경량 네트워크 구조와 자기지도(self-supervised) 학습 방식을 결합한 모델들이 등장하면서, 라벨링 비용을 줄이면서도 실시간 추론이 가능한 자기지도 단안 깊이 추정 모델이 활발히 연구되고 있다.

Lite-Mono는 CNN과 Transformer를 결합한 경량 단안 깊이 추정 모델로, 낮은 FLOPs 대비 우수한 성능을 보이는 것이 장점이다. 그러나 실제 구조를 분석하고 실험을 진행한 결과, 다음과 같은 한계점들이 드러났다.

(1) 공간적 관계 학습의 부족: Lite-Mono의 핵심 모듈인 LGFI(Local Global Feature Interaction) 블록은 연산량 절감을 위해 채널 차원에서만 attention 연산을 수행한다. 이로 인해, 물체의 형태·경계와 같이 공간 좌표 간의 관계를 직접적으로 반영하는 공간적인 주목(spatial attention)이 부족할 수 있다.

(2) 자기지도 학습 특유의 텍스처 부족 영역 문제: 기존 자기지도 단안 깊이 추정은 Pose Network를 통해 얻은 카메라 포즈 정보를 기반으로 Image Reconstruction Loss만을 사용해 학습한다. Ground Truth 없이 학습하기 때문에, 하늘이나 건물 벽처럼 텍스처가 거의 없는 영역에서는 깊이 추정이 불안정해지고, 잘못된 깊이가 할당되는 문제가 자주 관찰된다.

(3) Hard Regression 방식의 깊이 구간 가중치 문제: 기존 Lite-Mono는 연속적인 깊이 값을

직접 회귀(regression)하는 구조를 사용한다. 이 경우, 모든 깊이 구간을 동일한 중요도로 취급하게 되고, 실제로 많은 물체가 몰려 있는 깊이 범위(예: 차량, 보행자 등이 자주 존재하는 거리)에 대한 표현력이 부족하다.

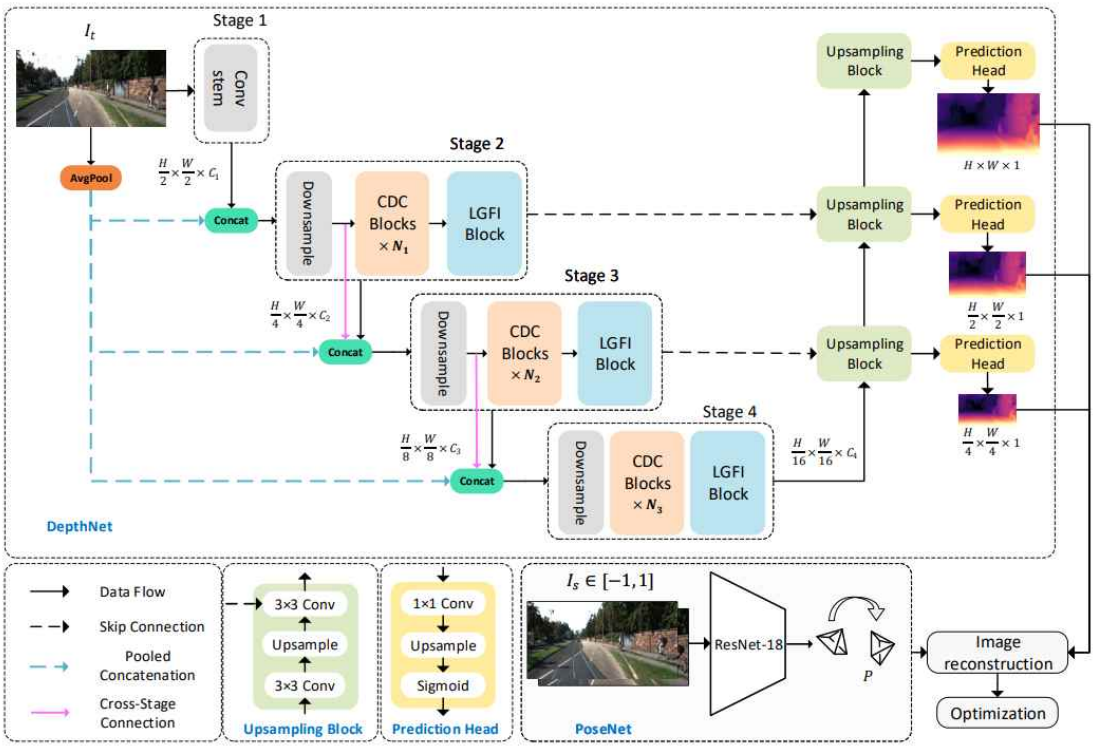
이러한 세 가지 문제의식을 바탕으로, 본 연구는 경량성을 유지하면서도 공간 정보 보완, 텍스처 부족 영역 개선, 깊이 구간 중요도 반영을 동시에 달성할 수 있는 Lite-Mono 개선 방안을 제안하고자 한다.

1.2 기존 기술

이미지로부터 CNN을 이용해 깊이를 회귀하는 단안 깊이 추정 기법이 발전해 왔으나, 대규모 정확한 깊이 GT를 얻기 어렵다는 한계 때문에, 최근에는 스테레오 영상이나 단안 비디오에서 이미지 재구성 손실을 이용해 학습하는 자기지도 단안 깊이 추정이 주류가 되었고, Monodepth 계열 방법들이 이 틀을 정립했다.

스테레오 기반 방법(Monodepth1)은 카메라 포즈 값이 이미 주어져 있기에, Depth Network만으로 학습할 수 있는 반면, 단안 비디오 기반 방법(Monodepth2)은 카메라 포즈를 함께 추정하기 위해 추가 Pose Network가 필요하다. 그럼에도 스테레오 데이터 수집·보정이 번거롭기 때문에, 실제 응용에서는 단안 비디오만으로 학습 가능한 자기지도 방식이 더 선호된다. 이후 연구들은 가려짐·이동 물체 문제를 줄이기 위한 손실 함수 개선과, 더 나은 네트워크 구조 설계에 집중해 왔다.

한편, 기존 CNN기반 네트워크들은 국소 receptive field에 기반하므로 장거리(global) 문맥 정보를 충분히 담기 어렵고, 이를 보완하기 위해 모델을 깊게 만들면 파라미터 수와 FLOPs가 크게 증가하는 문제가 있다. Vision Transformer(ViT)는 Self-Attention으로 전역 정보를 직접 모델링할 수 있지만, MHSA 연산량이 커서 경량·실시간 모델 설계에는 제약이 있다.



▲그림2. Lite-mono의 기본 구조

이러한 배경에서 제안된 Lite-Mono는 경량성과 정확도의 균형을 목표로 한 하이브리드 CNN–Transformer 기반 자기지도 단안 깊이 추정 모델이다. 인코더 각 단계에서 CDC(Consecutive Dilated Convolutions)로 다중 스케일 로컬 특징을 추출하고, LGFI(Local-Global Features Interaction) 모듈로 전역 문맥을 주입한다. 이때 연산량을 줄이기 위해 공간 차원이 아닌 채널 차원에서 cross-covariance attention을 계산하는 구조를 채택하여, 적은 파라미터와 낮은 FLOPs로도 KITTI 등에서 기존 대형 모델보다 경쟁력 있는 성능과 빠른 추론 속도를 달성한다. 본 연구에서는 이 Lite-Mono를 기준선으로 삼아 이후 장에서 제안하는 개선 기법들을 적용한다.

2. 제안하는 연구주제

2.1 제안하는 아이디어

본 연구의 핵심 아이디어는 경량 자기지도 단안 깊이 추정 모델인 Lite-Mono를 기준선으로 삼고, 다음 세 가지 측면에서 모델의 성능을 향상시키는 것이다.

(1) 공간 정보 보완을 위한 Spatial Attention 결합: Lite-Mono의 LGFI(Local-Global Feature Interaction) 모듈은 연산량을 줄이기 위해 채널 차원에서만 attention 연산을 수행하므로, 물체의 경계나 형태와 같이 공간 좌표 간 관계를 직접 다루는 능력이 제한적이다.

이에 본 연구에서는 인코더 하위 단계 특징 맵에 CBAM 기반 Spatial Attention 모듈을 결합하여, channel-wise로 정제된 특징에 대해 위치별 중요도 맵을 한 번 더 계산하고 곱해줌으로써, 중요한 영역에 더 큰 가중치를 부여하도록 한다. 이를 통해 큰 FLOPs 증가 없이 Lite-Mono의 공간 표현력을 보완하는 것을 목표로 한다.

(2) Teacher Network 기반 Pseudo Label을 이용한 자기지도 학습 보조: 순수 자기지도 학습은 텍스처가 부족한 영역(하늘, 평탄한 벽 등)에서 깊이가 불안정하게 추정되는 문제가 있다. 이를 완화하기 위해, 본 연구에서는 고성능 사전학습 모델(Depth Anything v2)을 Teacher로 사용하는 지식 증류 형태를 도입한다. Teacher 모델로 KITTI train split 전체에 대해 pseudo depth(또는 disparity)를 생성하고, 이를 정규화한 뒤 Lite-Mono의 예측 결과와 비교하는 보조 회귀 손실(auxiliary regression loss) 를 추가한다. 이렇게 하면 여전히 포토메트릭 재구성 손실을 중심으로 자기지도 학습을 유지하면서도, Teacher가 제공하는 정보를 활용해 특히 텍스처 부족 영역의 깊이 품질을 향상시킬 수 있다.

(3) AdaBins 아이디어를 변형한 Disparity Binning + Chamfer Distance Loss 도입: 기존 Lite-Mono는 연속적인 깊이 값을 직접 회귀하는 hard regression 방식을 사용한다. 이 경우, 실제로 물체가 많이 존재하는 깊이 구간과 거의 비어 있는 구간을 동일한 비중으로 다루게 되어, 깊이 분포 특성을 반영하기 어렵다는 한계가 있다. 본 연구에서는 AdaBins의 아이디어를 자기지도 환경에 맞게 변형하여, 깊이 대신 inverse depth(disparity) 를 대상으로 일정 개수(예: 64개)의 adaptive bin center를 예측하고, 픽셀별로 각 bin에 대한 분포를 예측한 뒤 bin center의 가중합으로 disparity를 복원하는 구조를 도입한다. 이때 Teacher로부터 얻은 disparity 분포와 Student가 예측한 bin center 분포 사이의 차이는, Scale-invariant Chamfer Distance Loss로 정의하여 최적화한다. 이를 통해 Teacher 분포를 따라가면서도, 사물이 밀집된 깊이 구간에 더 많은 표현력을 집중할 수 있도록 한다.

정리하면, 본 연구는

(1) Spatial Attention으로 Lite-Mono의 공간 표현력을 보완하고,

(2) Teacher 기반 pseudo label로 자기지도 학습의 약점을 보완하며,

(3) Disparity Binning + Chamfer Distance로 깊이 분포 자체를 더 잘 모사하도록 하는

세 가지 아이디어를 적용 및 비교함으로써, 경량성을 유지하면서도 깊이 추정 성능을 향상시키는 것을 목표로 한다.

2.2 최종 목표 및 기대효과

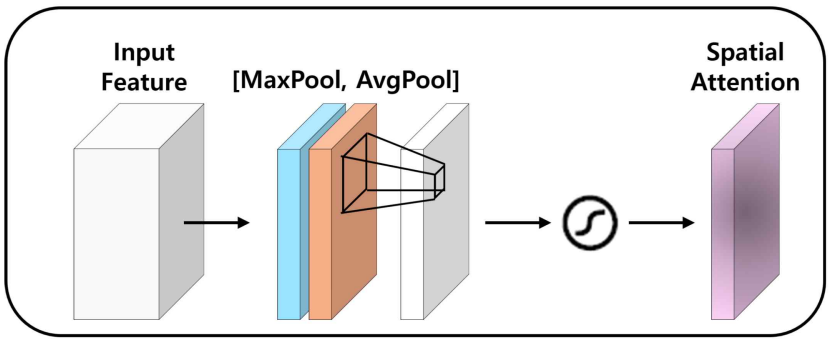
본 연구의 최종 목표는 경량 자기지도 단안 깊이 추정 모델인 Lite-Mono를 기반으로, 공간 정보를 보완하는 Spatial Attention, 고성능 Teacher 네트워크를 활용한 pseudo label 적용, 그리고 disparity binning과 Chamfer Distance Loss를 결합한 학습 방식을 통합함으로써, 기존 Lite-Mono와 유사한 연산량과 파라미터 수를 유지하면서도 KITTI와 같은 벤치마크에서 더 우수한 깊이 추정 성능을 달성하는 것이다. 즉, 모델의 구조적 복잡도를 크게 증가시키지 않으면서도 abs_rel, rmse 등의 정량적 지표와 시각적인 깊이 맵 품질을 함께 개선하는 것을 목표로 한다.

이를 통해 텍스처가 부족한 영역에서도 보다 안정적이고 자연스러운 깊이 맵을 얻을 수 있고, 물체가 많이 분포하는 깊이 구간에 대해서는 분포 기반 표현을 통해 기존 회귀 방식보다 향상된 표현력을 기대할 수 있다. 나아가 경량 CNN–Transformer 구조 위에 지식 증류와 binning 기법을 결합한 사례를 제시함으로써, 자율주행, 로봇, 증강현실 등과 같이 실시간 처리가 요구되는 임베디드 환경에서 활용 가능한 단안 깊이 추정 모델 설계에 하나의 구체적인 방향성을 제공하는 것을 기대한다.

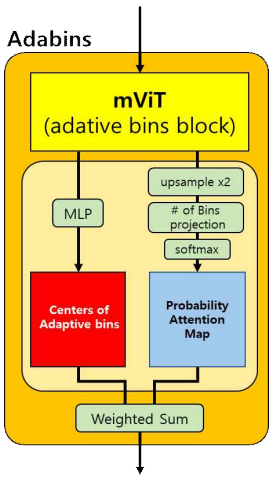
3. 본 연구와 관련된 이론 및 기술

3.1 본 연구에서 필요한 주요 이론 및 기술

CBAM(Convolutional Block Attention Module)은 합성곱 신경망에서 채널과 공간 두 관점에서 중요한 정보를 강조하기 위한 경량 어텐션 모듈이다. 먼저 채널 방향으로 평균 풀링과 최대 풀링을 수행해 채널별 통계값을 구한 뒤, MLP에 통과시켜 채널별 중요도 가중치를 계산하고 이를 원래 특징 맵에 곱해 채널을 정제(refine)한다. 이후 이렇게 정제된 특징에 대해 공간 방향으로 다시 평균 풀링과 최대 풀링을 수행하고, 이를 합성곱으로 처리해 위치별 중요도 맵을 얻은 뒤 원래 특징에 곱함으로써 공간 Attention을 수행한다. 이 과정에서 추가 파라미터와 연산량 증가는 매우 작으면서, 정보량이 많은 채널과 위치를 선택적으로 강조할 수 있다는 점이 특징이다.



▲그림3. Spatial Attention mdule (CBAM, ECCV2018)



▲그림4. AdaBins

AdaBins는 깊이를 하나의 연속 값으로 직접 회귀하는 대신, 깊이 축을 여러 개의 bin으로 나누고 bin center와 각 픽셀의 bin별 가중치를 함께 예측하는 방식의 metric depth 추정 기법이다. 네트워크는 전체 깊이 범위 내에서 다수의 bin center를 예측하고, 각 픽셀에 대해 bin별 확률 분포를 출력한 뒤, 이들의 가중합으로 최종 깊이 값을 계산한다. 이때 실제 데이터에서 물체가 많이 분포하는 깊이 구간에는 더 촘촘히 bin이 배치되므로, 전 범위를 균일하게 회귀하는 것보다 깊이 분포를 더 유연하고 정밀하게 표현할 수 있다. 원래 AdaBins는 실제 거리 값이 주어지는 지도학습 환경에서 metric depth를 예측하기 위해 제안되으며, 깊이를 스칼라가 아닌 분포와 구간 단위로 표현한다는 아이디어가 핵심이다.

3.2 제안하는 아이디어를 실현하기 위한 주요 이론 및 기술 분석

본 연구에서는 고성능 Teacher 모델이 가진 깊이 정보를 경량 Student 모델로 전달하기 위해 지식 증류(Knowledge Distillation)를 사용하고, 이를 bin 기반 표현과 결합하기 위해 Chamfer Distance를 distillation 손실로 활용한다.

지식 증류는 성능이 우수한 큰 모델(Teacher)이 가진 지식을 더 작은 모델(Student)에 전달해, Student가 적은 파라미터로도 Teacher에 가까운 성능을 내도록 하는 학습 기법이다. 분류 문제에서는 Teacher의 softmax 출력이나 중간 feature를 Student가 모방하도록 할 수 있으며, 깊이 추정과 같은 회귀 문제에서는 Teacher가 예측한 깊이 또는 disparity를 pseudo label로 간주해 Student 예측과의 L1 또는 L2 손실을 줄이는 방식이 사용 될 수 있다. 특히 자기지도 단안 깊이 추정에서는 포토메트릭 재구성 손실만으로는 텍스처가 부족한 영역이나 가려짐 영역에서 추정이 불안정해지기 쉬운데, Teacher가 제공하는 상대적으로 정확한 정보를 보조 정보로 사용하면 이러한 약점을 보완할 수 있다.

Chamfer Distance는 이러한 지식 증류를 AdaBins 기반의 bin 표현과 통합하기 위해 사용된다. Chamfer Distance는 두 점 집합 사이의 분포 차이를 측정하는 데 사용되는 거리 척도로, 한 점 집합의 각 점에 대해 다른 점 집합에서 가장 가까운 점까지의 거리를 계산하고 이를 평균 또는 합으로 취해 한 방향의 거리를 정의한 뒤, 반대 방향에 대해서도 동일한 값을 계산해 두 값을 합치는 방식이 일반적이다. 이 거리는 점의 순서에 영향을 받지 않고, 두 집합이 공간상에서 얼마나 잘 겹치는지, 즉 분포 형태가 얼마나 유사한지를 나타내는 데 적합하다.

$$d(X,Y) = \Sigma_{x\in X} \min_{y\in Y} |x-y|^2 + \Sigma_{y\in Y} \min_{x\in X} |y-x|^2$$

▲식1. Chamfer Distance

본 연구에서는 Teacher가 만드는 disparity 분포를 하나의 점 집합으로, 모델이 예측한 bin center 분포를 다른 점 집합으로 간주한 뒤, 이 둘 사이의 Chamfer Distance를 최소화하도록 학습함으로써, 명시적인 1:1 대응이 없는 상황에서도 Student의 bin들이 Teacher 분포를 따라가도록 유도한다. 즉 Chamfer Distance는 Teacher의 연속 disparity를 AdaBins 스타일 bin 표현과 연결해 주는 분포 기반 지식 증류 손실로 기능하며, 제안하는 아이디어를 실제로 구현하는 핵심 요소가 된다.

4. 연구 내용

4.1 제안하는 아이디어의 원리

본 연구의 전체적인 개선 흐름은 기존 Lite-Mono의 구조를 그대로 유지한 상태에서, 필요한 부분에만 최소한의 모듈과 손실을 추가해 성능을 향상시키는 것이다.

이를 위해 먼저 채널 차원 어텐션이 이미 포함되어 있는 Lite-Mono의 구조적 특성을 고려하여, CBAM 모듈에서 채널 어텐션은 제외하고 공간 어텐션만을 선택적으로 도입하였다. 기존 Lite-Mono의 LGFI 모듈이 채널 상관관계를 충분히 모델링하고 있기 때문에, 채널에 대한 어텐션을 또 추가하는 것은 중복이라는 판단에서였고, 대신 LGFI 이후의 특징 맵에 대해 Spatial Attention만 적용하여 깊이 추정에 중요한 위치 정보만 한 번 더 강조하는 방향으로 설계하였다. 즉, 채널 정보는 Lite-Mono에 맡기고, 본 연구는 공간 정보만 보완하는 것을 원리로 삼았다.

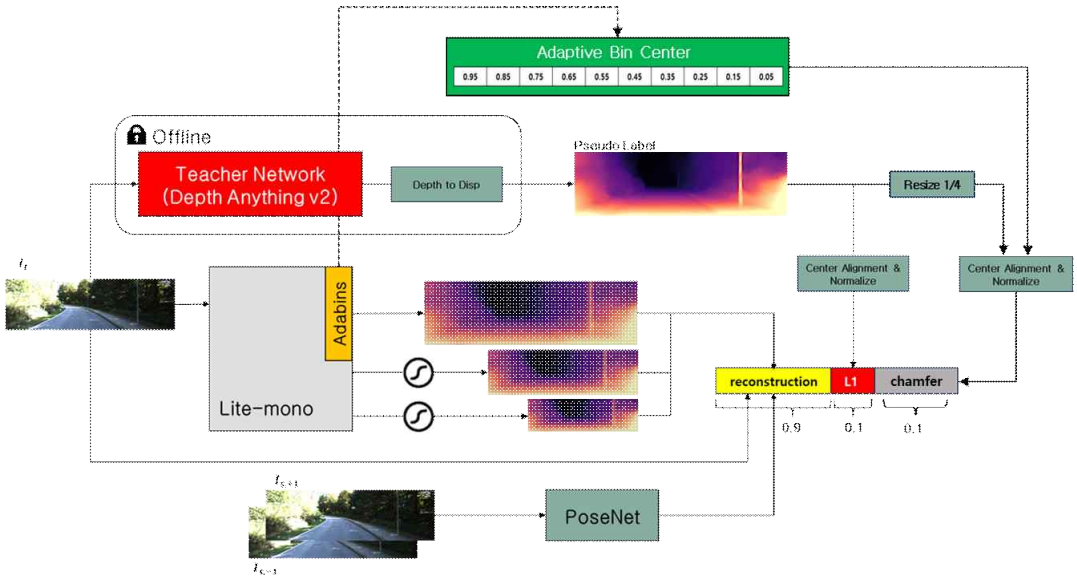
두 번째로, 지식 증류를 활용하면, 텍스처 부족 문제를 해결할 수 있는데, Depth Anything V2와 같은 대형 Teacher 모델은 다양한 데이터셋으로 학습되어 하늘과 같이 텍스처가 부족한 영역에서도 구조적인 깊이 패턴을 잘 학습할 수 있다. 따라서 Teacher의 출력을 pseudo disparity로 가져와 Student를 보조 감독해 준다면, 자기지도 손실이 "놓치는 부분"을 Teacher가 채워줄 수 있고, 특히 텍스처 부족 영역이나 occlusion 주변에서 깊이가 더 안정적으로 수렴할 것이라 예상했다. 다만 self-supervised 환경에서는 절대 스케일이 중요하지 않으므로, Teacher 값의 절대 크기보다 중간값 기준 정렬(median center alignment) 이후의 상대적인 분포만 맞추는 것이 더 적절하다고 보았다. 이렇게 하면 Teacher의 구조 정보는 가져오되, 서로 다른 데이터셋·학습 설정에서 비롯된 스케일 차이는 크게 문제되지 않게 된다.

세 번째로, Binning 전략을 활용하는 것의 장점은, 깊이를 단순히 한 점으로 회귀하는 것보다 "어떤 깊이 구간에 얼마나 많이 분포하는지"라는 분포 정보를 함께 배우는 것이 더 풍부한 표현력을 줄 수 있다는 것이다. 실제 장면에서는 물체가 특정 깊이 범위(예: 차량, 건물, 보행자 거리)에 많이 몰려 있고, 매우 먼 영역(하늘, 원거리 배경)은 상대적으로 정보량이 적다. 그런데 기존 단순 회귀는 깊이 축 전체를 균일하게 다루어 이런 특성을 살리지 못한다. 그래서 AdaBins처럼 disparity 축을 여러 bin으로 나누고, 그 위치와 분포를 학습하면, 모델이 실제로 자주 등장하는 거리 구간에 더 많은 표현력을 할당할 수 있을 것이라고 기대했다. 이때 Teacher의 disparity 분포를 기준으로 Student의 bin center를 끌어당기면, bin들이 데이터/Teacher가 자주 사용하는 영역에 자동으로 몰리게 되어, 단순 균일 bin보다 효율적으로 깊이 축을 쓸 수 있다.

4.2 전체 기능도

전체 시스템의 기능 흐름은 오프라인 pseudo label 생성 단계와 온라인 학습 단계로 나뉜다. 오프라인 단계에서는 먼저 Depth Anything V2에 전체 학습 데이터셋을 입력하여 각 프레임에 대한 pseudo disparity를 계산하고, 이를 프레임 경로 또는 인덱스를 키로 하여 저장한다. 이때 이후 분포 기반 학습에 활용하기 위해 disparity의 전체 범위와 분포에 대한 기본 통계도 함께 분석해 둔다.

온라인 학습 단계에서 데이터로더는 단안 카메라의 시퀀스데이터로부터 타깃 프레임과 인접 프레임들을 로딩하고, 동시에 타깃 프레임에 대응하는 pseudo disparity를 함께 불러온다. 타깃 프레임은 CBAM Spatial Attention이 삽입된 Lite-Mono 기반 Student 네트워크로



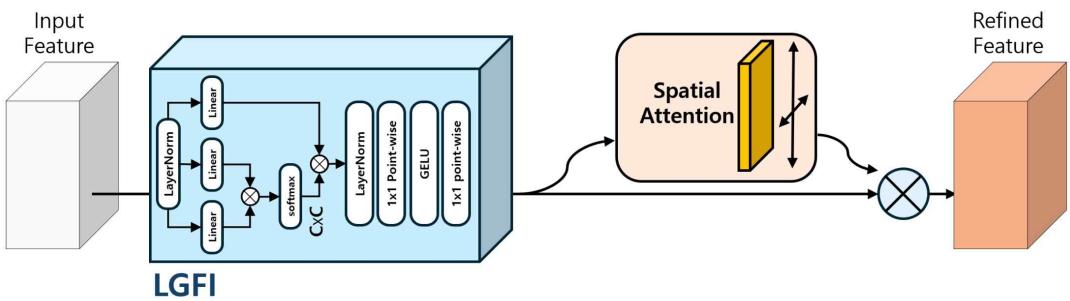
▲그림5. 전체 구조도

전달되어 다중 스케일 disparity와, 최상위 스케일에서 disparity bin center 및 픽셀별 bin 가중치를 출력한다. Student가 예측한 disparity는 인접 프레임들과 PoseNet이 예측한 상대 포즈와 함께 사용되어 소스 프레임을 타깃 뷰로 재투영하는 데 쓰이고, 여기서 photometric 재구성 손실과 smoothness 손실이 계산된다.

동시에 데이터로더가 가져온 Teacher pseudo disparity는 median center alignment와 정규화를 거친 뒤 Student의 disparity와 비교되어 픽셀 단위 지식 증류 손실이 계산된다. 분포 기반 지식 증류를 위해서는 Teacher disparity를 일정 규칙에 따라 샘플링하여 1차원 점 집합 형태로 변환하고, Student가 예측한 bin center를 다른 점 집합으로 두어 Chamfer Distance를 계산한다. 최종적으로 자기지도 손실, 픽셀 단위 distillation 손실, Chamfer Distance 기반 분포 손실을 가중합한 값이 역전파를 통해 Student와 PoseNet의 파라미터를 갱신한다.

4.3 주요 기능의 구현 방법

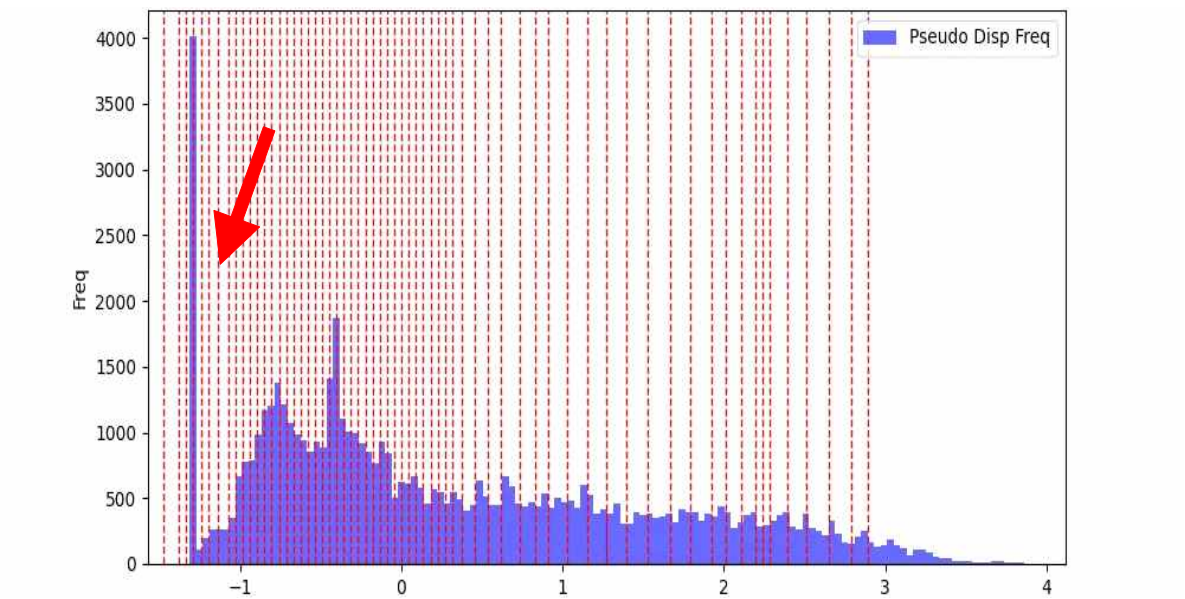
CBAM 모듈 적용 과정에서는 CBAM 구조에서 Spatial Attention 블록만 따로 분리하여 모듈화하고, 인코더 상단의 고수준 특징 맵을 입력으로 받아 동일한 크기의 Attention 맵을 출력하도록 하였다. 이 Attention 맵은 시그모이드를 거쳐 0~1 범위로 정규화된 뒤 원래 특징 맵에 곱해지는 방식으로 적용된다.



▲그림6. Spatial Attention module의 삽입

지식 증류를 위한 pseudo label 생성은 학습과 분리된 전처리 스크립트로 구현하였다. Depth Anything V2의 프리트레인 모델을 불러와 KITTI 학습 split의 모든 이미지를 순회하면서 disparity를 추론하고, 결과를 압축된 텐서 파일로 저장하였다. 이후 데이터로더는 이미지 파일 경로를 기준으로 동일한 이름 또는 인덱스를 갖는 pseudo disparity 파일을 함께 로딩하도록 수정하였다. 학습 루프에서는 로딩된 Teacher disparity에 대해 먼저 median 값을 기준으로 center alignment를 수행하고, 전체 범위를 0~1 사이로 정규화한 뒤 Student의 예측과 L1 손실을 계산하였다. 이렇게 함으로써 Teacher와 Student의 출력 스케일 불일치 문제를 해결하고, self-supervised 학습에서 일반적으로 사용하는 disparity 범위와도 일관성을 맞추었다.

Chamfer Distance와 disparity binning 적용은 Student 디코더에 bin 예측 분기를 추가하는 방식으로 구현하였다. 최상위 스케일의 디코더 출력에서 하나의 헤드는 K개의 bin center를, 다른 헤드는 각 픽셀에 대한 K차원 가중치 맵을 예측하도록 설정하고, 가중치 맵에는 softmax를 적용해 픽셀별 분포로 만들었다. 최종 disparity는 이 분포와 bin center의 가중합으로 계산되며, 학습 시에는 Teacher disparity에서 0이 아닌 값들을 일정 간격으로 샘플링하여 Teacher 점 집합을 구성하고, Student bin center를 다른 점 집합으로 두어 Chamfer Distance를 계산하였다. 이때 disparity binning을 적용하면 하늘과 같이 깊이가 사실상 무한대인 영역이 disparity 0으로 몰리면서 0 값에 대한 쏠림 현상(그림7)이 발생하므로, 학습 시 disparity가 0인 픽셀을 마스킹하여 Chamfer Distance와 distillation 손실 계산에서 제외하는 방식으로 이 문제를 해결하였다. 이렇게 구성된 Chamfer Distance 손실은 하이퍼파라미터를 통해 전체 손실에 적절히 가중하여 포함하였고, CBAM만 적용한 모델, CBAM + 픽셀 단위 지식 증류 모델, CBAM + 지식 증류 + disparity binning 및 Chamfer Distance 모델을 순차적으로 학습하면서 각 단계의 구현이 성능과 학습 안정성에 미치는 영향을 비교·검증하였다.



▲그림7. disparity의 0값 쏠림 현상

5. 연구 결과

5.1 최종 결과

본 연구에서는 KITTI eigen_zhou split을 기준으로 Lite-Mono 원 논문 결과, 동일 설정으로 재학습한 baseline, Teacher 기반 지식 증류만 적용한 모델, 고정 bin을 사용한 disparity binning 모델, 그리고 제안하는 Adaptive Disparity Binning + Chamfer Distance 모델을 비교하였다.

먼저 Lite-Mono를 논문 설정과 동일하게 재현한 baseline은 FLOPs 5.03G 수준에서 원 논문 결과와 유사한 지표(abs_rel 0.107, rmse 4.63, rmse_log 0.184, δ1 0.888)를 얻어, 구현과 학습 설정이 적절히 재현되었음을 확인하였다. 여기에 Spatial Attention(CBAM의 공간 부분만)을 추가한 경우 FLOPs 증가 없이 소폭의 지표 개선이 있었으나, 전체 틀에서 큰 변화를 줄 정도의 차이는 아니었으므로 이후 결과 표에서는 최종 모델에 포함된 형태로만 보고하였다.

Teacher(Depth Anything V2)를 이용해 pseudo disparity를 생성하고, 픽셀 단위 L1 distillation loss만 추가한 모델은 abs_rel은 baseline과 거의 동일한 수준을 유지하면서 sq_rel, rmse, rmse_log, δ2, δ3가 전반적으로 개선되는 결과를 보였다. 특히 텍스처가 부족한 영역에서 깊이 맵이 더 안정적으로 추정되며, 경계 부근의 노이즈가 감소하는 시각적 개선을 확인할 수 있었다. 다만 Teacher 출력을 정규화하여 사용한 영향으로, 절대 정확도를 나타내는 δ1 지표는 소폭 감소하는 trade-off를 보였다.

고정 bin을 사용한 disparity binning 모델은 FLOPs를 약간 증가시키면서도 성능 향상은 거의 없거나 일부 지표에서 오히려 악화되는 경향을 보였다(abs_rel 및 rmse 소폭 악화). 이는 bin 경계가 데이터 분포나 Teacher 분포를 고려하지 않고 균일하게 설정되어, 실제 물체가 많이 존재하는 거리 구간에 표현력을 집중하지 못한 결과로 해석된다.

반면, 제안하는 Adaptive Disparity Binning + Chamfer Distance 모델은 FLOPs가 약 8.68G로 증가했음에도 abs_rel 0.106, sq_rel 0.762, rmse 4.47, rmse_log 0.175, δ2 0.967, δ3 0.986 등 거의 모든 지표에서 baseline과 논문 보고값을 상회하는 성능을 보였다. 특히 rmse와 rmse_log 지표 개선이 두드러져, 전체 깊이 분포 관점에서 오차가 안정적으로 감소했음을 확인할 수 있었다. 시각적으로도 제안 모델은 하늘·건물·도로 등 넓은 영역에서의 깊이 변화가 더 부드럽고, 물체 경계가 상대적으로 깔끔하게 표현되는 결과를 보여주었다.

종합하면, 지식 증류만 적용한 모델과 Adaptive Disparity Binning + Chamfer Distance 모델이 baseline 대비 가장 일관된 성능 향상을 보였으며, 이 중 최종 제안 모델이 정량·정성 평가 모두에서 가장 우수한 결과를 달성하였다.

Model	Lower is better					Higher is better		
	FLOPs	abs_rel	sq_rel	rmse	rmse_log	a1	a2	a3
Lite-Mono - Paper (640x192)	**5.032G**	0.107	0.765	4.561	0.183	0.886	0.963	0.983
Lite-mono - Trained	**5.032G**	0.107	0.834	4.631	0.184	**0.888**	0.963	0.982
LiteMono + Adabins (Chamfer dist loss)	8.675G	**0.106**	**0.762**	**4.474**	**0.175**	0.886	**0.967**	**0.986**
Pseudo Label	**5.032G**	0.107	0.766	4.494	0.177	0.883	0.966	0.985
Fixed bin	5.694G	0.109	0.864	4.669	0.186	0.888	0.962	0.982
Spatial Att	5.033G	**0.106**	0.796	4.589	0.183	**0.889**	0.963	0.982

▲표1. 학습결과 정리

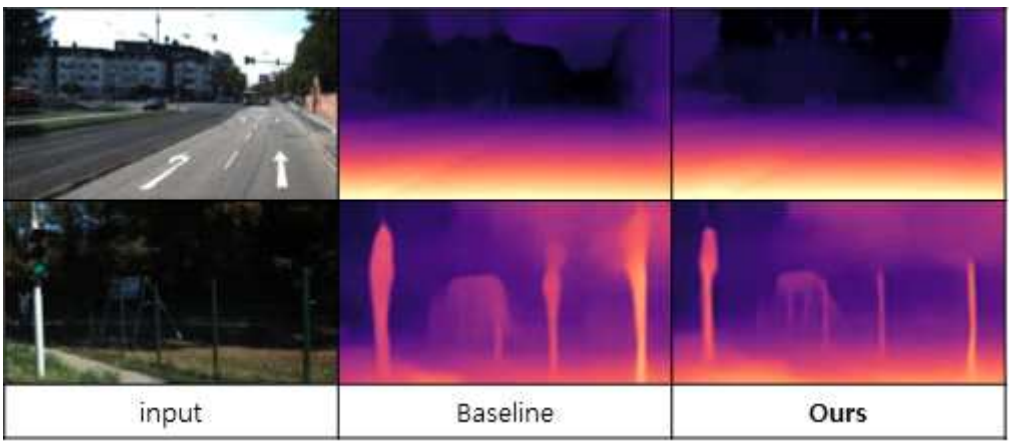
5.2 결과 분석

실험 결과를 바탕으로 각 단계별 아이디어가 성능에 미친 영향을 분석하면 다음과 같다. 먼저 Spatial Attention(CBAM의 공간 부분)만을 Lite-Mono에 추가한 실험에서, FLOPs 변화 없이 일부 지표가 소폭 개선되었기는 했으나, 여전히 공간 위치별 중요도를 정확히 반영하지는 못한다고 판단된다. 다만 이 효과는 경량 구조에 최소한의 모듈을 추가한 수준이기 때문에, 단독으로는 성능을 크게 끌어올리지는 못했다고 해석할 수 있다.

Teacher 기반 지식 증류를 픽셀 단위 L1 손실로만 적용한 경우, rmse와 rmse_log와 같은 분포 기반 지표가 개선된 반면 δ1이 약간 감소한 것은, 정규화된 Teacher disparity를 사용하면서 절대 스케일보다는 상대적인 구조를 따라가도록 학습되었기 때문으로 볼 수 있다. self-supervised 설정에서는 어차피 절대 스케일이 잘 정의되지 않기 때문에, Teacher의 구조 정보를 받아들이는 과정에서 δ1 지표가 약간 손해를 보는 대신, 전체적인 깊이 맵의 모양과 구조가 더 안정화되는 방향으로 수렴한 것으로 해석된다.

고정 bin 기반 disparity binning은 오히려 일부 지표가 악화되는 결과를 보였는데, 이는 bin 경계를 데이터나 Teacher 분포와 무관하게 균일하게 나눈 설계의 한계로 볼 수 있다. 실제로 Teacher disparity 분포를 분석해 보면 특정 거리 구간에 샘플이 편중되어 있으며, 균일 bin은 이런 분포를 반영하지 못한다. 따라서 고정 bin 구조에서는 오히려 표현력이 분산되어, Lite-Mono의 단순 회귀보다도 유리한 점이 뚜렷하지 않았던 것으로 해석된다.

이에 비해 Adaptive Disparity Binning + Chamfer Distance 모델은 Teacher 분포를 기준으로 bin center를 학습하도록 설계했기 때문에, Teacher가 자주 사용하는 disparity 구간 근처에 bin들이 밀집하게 되고, 결과적으로 물체가 많이 존재하는 깊이 범위에 표현력을 집중할 수 있었다. median center alignment와 정규화를 통해 Teacher와 Student 사이의 스케일 차이를 줄이고, Chamfer Distance를 사용해 두 분포의 모양을 직접 맞추도록 한 것이 rmse와 rmse_log의 안정적인 개선으로 이어진 것으로 볼 수 있다. 특히 disparity가 0인 무한 깊이 영역을 학습 시 마스킹하여 Chamfer Distance와 distillation 손실 계산에서 제외한 부분은, 하늘 영역에 대한 과도한 쏠림을 방지하는 데 중요한 역할을 하였다. 만약 이 마스크 없이 학습을 진행하면, bin center들이 0 근처에 과도하게 몰리면서 실질적으로 관심 있는 거리 범위에 대한 해상도가 떨어지는 문제가 발생한다.



▲그림8. 결과 시각화

제안된 방법으로 얻은 깊이 맵(그림7)은 전반적으로 Baseline에 비해 구조가 더 또렷하고 안정적으로 표현되었다. 먼저 기둥과 같은 얇은 구조물을 보면, Baseline에서는 폭이 두껍게 번지거나 주변 배경과 섞여 형체가 흐려지는 반면, 제안 모델의 결과에서는 기둥이 위아래

로 가늘고 연속적으로 유지되며 실제 물체의 형태가 뚜렷하게 드러난다. 또한 하늘과 같은 원거리 영역에서 Baseline은 깊이값이 오검출 되는 것으로 보이는 반면, 제안된 결과는 색상이 비교적 균일하게 유지되어 원거리 영역의 깊이가 더 안정적으로 추정된 것으로 해석할 수 있다.

6. 연구 과정 중 발생한 문제점 및 해결 방법

본 연구를 진행하는 과정에서 여러 가지 기술적 문제들이 발생하였으며, 각 문제에 대해 실험을 통해 원인을 분석하고 해결 방안을 적용하였다. 주요 문제와 해결 방법은 다음과 같다.

첫째, Teacher 네트워크(Depth Anything V2)와 Student(Lite-Mono) 사이의 스케일 불일치 문제가 발생하였다. Depth Anything V2는 약 0~150 정도의 범위를 갖는 disparity 값을 출력하는 반면, Lite-Mono는 0~1 구간의 정규화된 disparity를 회귀하도록 설계되어 있어, 두 출력을 그대로 L1 loss로 비교하면 distillation이 제대로 이루어지지 않았다. 이를 해결하기 위해 각 프레임별 Teacher disparity에 대해 median center alignment를 적용한 후 정규화하는 전략을 도입하였고, 이를 통해 Teacher–Student 간 스케일 차이를 완화하고 distillation loss가 안정적으로 작동하도록 만들었다.

둘째, disparity binning을 적용하는 과정에서 0 값 쏠림 문제가 발생하였다. 하늘과 같이 깊이가 사실상 무한대인 영역은 disparity가 0에 해당하므로, Teacher disparity 분포를 그대로 사용할 경우 0 값 주변에 데이터가 과도하게 몰리는 현상이 나타났다. 이 상태에서 Chamfer Distance를 계산하면 Student의 bin center들이 0 근처로만 끌려가 실제 관심이 있는 거리 구간에 대한 표현력이 떨어지는 문제가 발생하였다. 이를 해결하기 위해 disparity가 0인 픽셀을 마스크 처리하여 Chamfer Distance 및 distillation 손실 계산에서 제외하도록 변경하였고, 그 결과 bin center들이 실제 물체가 존재하는 disparity 구간에 더 잘 분포하도록 학습이 이루어졌다.

셋째, AdaBins 구조를 그대로 차용하려 할 때 self-supervised 환경과의 부적합 문제가 드러났다. 원래 AdaBins는 metric depth를 직접 사용하는 지도학습 환경을 가정하고 bin을 설계하기 때문에, disparity(0~1)를 사용하는 self-supervised Lite-Mono에 직접 적용하면 bin 경계와 분포가 실제 학습 설정과 맞지 않았다. 초기에는 균일한 고정 bin을 disparity 축에 그대로 적용해 보았으나, 성능이 거의 향상되지 않거나 일부 지표가 악화되는 결과가 나타났다. 이에 따라 실제 metric depth 기준이 아닌 Teacher disparity 분포와 Student bin center 분포를 Chamfer Distance로 정렬하는 방식을 채택하여, bin 경계가 데이터 분포에 따라 자동으로 형성되도록 수정하였다. 이로써 self-supervised 설정에서도 binning의 이점을 살릴 수 있었다.

마지막으로, 지식 증류를 온라인 방식으로 구현할 경우 학습 속도 및 자원 사용량 문제가 있었다. 초기 구현에서는 매 학습 iteration마다 Teacher(Depth Anything V2)를 함께 추론하여 Student를 감독하는 구조를 고려하였으나, 모델 규모와 KITTI 데이터셋의 크기를 감안했을 때 학습 시간이 지나치게 길어지고 GPU 메모리 사용량도 크게 증가하는 문제가 발생할 것으로 예상되었다. 이를 피하기 위해 Teacher 추론은 학습 루프 밖에서 한 번만 수행하고, 전체 데이터셋에 대한 pseudo disparity를 오프라인으로 사전 생성·저장한 뒤 데이터로더에서 함께 로딩하는 방식으로 변경하였다. 이로써 지식 증류를 적용하면서도 학습 속도를 크게 저하시키지 않고 실험을 진행할 수 있었다.

7. 팀원 역할 분담

이름	역할	비고
한승민	아이디어 제안, 모델 개선 및 실험 설계, 실험 진행 및 결과 확인	팀장

8. 소감

본 연구를 수행하면서 단안 깊이 추정 문제는 단순히 "깊이 값을 잘 맞추는 회귀 문제"가 아니라, 데이터 분포, 스케일, 표현 방식, 학습 손실 설계가 서로 강하게 얽혀 있는 문제라는 것을 체감할 수 있었다. 특히 Lite-Mono 구조를 그대로 재현해 보고, 그 위에 CBAM, 지식 증류, binning, Chamfer Distance를 하나씩 쌓아가면서 "이론적으로는 좋아 보이는 아이디어도 실제 코드와 학습 환경에 맞게 조정하지 않으면 오히려 성능을 해칠 수 있다"는 점을 여러 번 경험했다.

또한 Depth Anything V2와 같은 대형 사전학습 모델을 Teacher로 활용해 보면서, 단순히 더 큰 모델을 가져다 쓰는 것을 넘어, 그 모델이 가진 정보를 어떤 형태(픽셀 값, 분포, 스케일)를 통해 어떻게 전달할지 설계하는 과정이 중요하다는 것을 배웠다. 특히 disparity 스케일 정렬, 0 값 마스킹, Chamfer Distance와 같은 세부 설계들이 최종 성능에 큰 영향을 준다는 점이 인상적이었다.

마지막으로, 실험 과정에서 여러 실패 사례와 예상과 다른 결과들을 반복해서 마주쳤지만, 하나씩 원인을 추적하고 해결해 나가는 과정을 통해 논문에서 본 아이디어를 실제 구현과 실험으로 옮기는 능력을 키우는 데 큰 도움이 되었다. 이번 졸업 연구를 통해 얻은 경험을 바탕으로, 앞으로도 자율주행이나 3D 인지와 관련된 연구를 더 깊이 있게 이어가고 싶다는 생각을 하게 되었다.

9. 참고문헌

[1] Clement Godard, Oisin Mac Aodha, Michael Firman, and ´Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In ICCV, 2019.

[2] Ning Zhang, Francesco Nex, and George Vosselman. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In CVPR 2023.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth Estimation Using Adaptive Bins. In CVPR 2021.

[4] Sanghyun Woo , Jongchan Park , Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In ECCV 2018.

[5] Yifan Liu, Changyong Shun, Jingdong Wang, and Chunhua Shen. Structured Knowledge Distillation for Dense Prediction. In CVPR 2019.