https://doi.org/10.7840/kics.2023.48.12.1623

자율주행을 위한 Monocular Depth Estimation에 관한 연구

유 지 상◆, 전 우 민*, 이 성 진°

Research on Monocular Depth Estimation for Autonomous Driving

Jisang Yoo◆, Woomin Jun*, Sungjin Lee°

요 약

자율주행을 위해서는 주변상황을 정확히 인지하기 위해 다양한 센서들, 즉 카메라, 라이다, 레이다 등으로부터 정보를 획득하여 객체인식, 주행영역 인식, 차선인식, 거리예측 등의 상황인지 작업들을 수행해야 한다. 하지만, 이런 여러 센서들로부터의 상황인지 연산은 상당한 고비용, 고연산 및 고지연을 요구하며 이는 실시간으로 엣지컴퓨팅을 수행해야 하는 자율주행 시스템에서 현실적 구현의 어려움을 촉발한다. 이에 특히 3차원의 방대한 포인트 클라우드 데이터를 지니는 라이다 혹은 레이다 센서를 사용하지 않고 카메라 만을 사용하여 상황인지를 수행하는 연구가 주요하게 수행되고 있다. 본 연구에서는 하나의 카메라 만을 이용하여 주변 상황의 3차원 정보를 얻어내는 MDE (Monocular Depth Estimation) 방식의 성능 최적화 방안에 대해 연구하였다. 특히 고전적 데이터 증식 방식과 제안하는 합성기반 데이터 증식 방식들을 사용하여 정확도를 올리는 방식에 대해 알아보았다. 실험 결과 제안하는 데이터 증식 방식과 최적 손실함수를 사용하였을 경우 REL를 약 3.9% 줄일 수 있었다.

키워드 : 자율주행, 단안 깊이 예측, 딥러닝, 데이터 증강
Key Words : Autonomous Driving, Monocular Depth Estimation, Deep Learning, Data Augmentation

ABSTRACT

To achieve autonomous driving, various sensors such as cameras, lidar, and radar are used to accurately perceive the surrounding environment. However, processing information from these multiple sensors for situational awareness requires significant costs, computational power, and introduces high latency. This poses practical challenges for real-time implementation in autonomous driving systems that require on-the-edge computing. Especially, research is actively underway to perform situation awareness using only cameras, without utilizing lidar or radar sensors that generate extensive 3D point cloud data. In this study, we explored ways to optimize the performance of Monocular Depth Estimation, a method that derives 3D information of the surrounding environment using only a single camera. We particularly focused on optimizing this approach by utilizing classical data augmentation techniques, proposing synthetic data augmentation methods, and employing appropriate loss functions to achieve the best results.

※ 이 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2023 신산업특화선도전문대학 지원사업)
◆ First Author : Dong-Seoul University, Department of Electric Engineering, jisang1103@naver.com, 학생회원
° Corresponding Author : Dong Seoul University Department of Electronic Engineering, sungjinlee@du.ac.kr, 정회원
* Dong-Seoul University, Department of Electric Engineering, aplus912@naver.com, 학생회원
논문번호 : 202308-057-C-RU, Received August 23, 2023; Revised September 8, 2023; Accepted September 8, 2023

Ⅰ. 서 론

자율주행 기술은 그 활용성이 매우 다양하고 산업전체에 미치는 영향이 지대하여 최근 가장 많이 연구되고 있는 기술 분야 중 하나이다. 이런 자율주행 차량들은 실제 운전자 없이도 주변 환경을 인식하고 분석하여 목표 위치까지 안전하게 운행하는 능력을 갖추고 있다. 그러나 이러한 혁신적 기술이 여전히 직면하고 있는 핵심 문제 중 하나는, 차량 주행 중 3차원 공간을 정확하고 빠르게 이해하고 처리하는 능력이다[1,2].

3차원 공간 인식은 자율주행 시스템에서 중요한 역할을 담당하며, 이를 효과적으로 수행하고 상업적으로 널리 활용되기 위해서는 단일 카메라 만을 이용하여 빠른 시간 안에 적은 비용으로 객체와 장면에 대한 깊이예측 (Depth Estimation)을 수행하는 것이 선호된다[3,4]. 이 단안깊이예측 (MDE: Monocular Depth Estimation)은 차량 주변 환경을 이해하여 장애물을 피하고, 안전한 경로를 계획하는 데 필요한 3차원 정보를 제공함으로서 자율주행차의 안전성과 효율성을 극대화하는 기술이다[3,4].

현재 단안깊이예측 기술분야에는 다양한 기법들이 개발되어 왔다[3-5]. 그러나 이들은 종종 복잡한 조명 상황, 다양한 기상 조건, 또는 물체와 텍스처의 다양성과 같은 실제 세계의 변동성에 대응하는 데 한계를 가지고 있다.

하지만, 단안깊이예측 모델의 정확도는 여전히 크게 개선될 여지가 있으며, 이는 앞서 언급한 복잡한 조명 상황, 다양한 기상 조건, 그리고 물체와 텍스처의 다양성 등과 같은 실제 세계에 대한 적절한 훈련 데이터의 다양성을 증가시킨다면 해결될 수 있을 것이다. 본 연구에서는, 이런 다양한 데이터 증강 기법을 통해 단안깊이예측 모델의 성능 향상에 어떻게 영향을 미칠 수 있을지에 대해 분석하고 이에 대한 방법론을 제시한다. 즉, 다양한 데이터 증강 기법과 손실함수들을 적용하고, 이러한 기법이 모델의 정확도에 어떤 영향을 미치는지를 연구하였다. 또한, 단안깊이예측를 위한 최적의 데이터 증강 전략을 탐색하고 이를 통해 기존 모델의 성능을 개선하는 방법을 제안한다.

Ⅱ. 관련 연구

딥러닝 출현 이전의 초기 단안깊이예측 연구는 주로 Depth-Cue 기반으로 연구되었었다[6-8]. 연구 [6]에서는 Vanishing Point에 기반하여 접근하였으며, 연구 [7]에서는 focus와 defocus에 기반하여 접근하였고 연구 [8]에서는 shadow기반으로 접근하였다. 하지만 이런 연구들은 제한된 조건하에서 단안깊이예측을 할 수 있는 약점이 존재하여 범용으로 사용할 수 있는 기술은 아니었다.

딥러닝의 출현과 발전으로[9-11], 이런 단안깊이예측 분야에서도 딥러닝을 활용한 연구가 시작되었다[12]. 이는 인코더-디코더 구조에 기인하여 RGB 입력을 받아 깊이맵을 도출하는 형태이다. 이런 인코더-디코더 구조와 유사하게 수많은 연구들[12-19]이 이후에 출현하게 되었다. 이에 더해, 연구 [20-22]에서는 인코더의 출력 특성맵에 CRF (Conditional Random Field)를 통한 연속된 영상들의 확률적 결합에 기반하여 깊이맵을 생성하는 연구가 진행되었다. 연구 [20]에서는 연속된 영상들을 여러 크기의 특성맵으로 추출한 뒤 이를 어텐션 기반으로 결합하여 깊이맵으로 도출하였다. 또한 이런 CRF의 적용방식을 다양화 하여 연구 [23]에서는 multiple cascade CRF, 연구 [21]에서는 continuous CRF, 연구 [22]에서는 hierarchical CRF, 연구 [24]에서는 FC-CRF 방식을 통해 단안깊이예측을 수행하였다.

하지만, 단안깊이예측을 지도학습에 적용하기 위해서는 높은 데이터 레이블링 비용이 요구되기 때문에 이 비용을 낮추기 위해 비지도학습 기반으로 접근하여 해결하려는 시도가 있었다[25-30]. 또한, 이런 비지도 학습에 더해 기존의 데이터를 데이터 증식, 스타일 변환, 데이터 합성과 같은 방법으로 최대한 증식하여 활용하려는 시도 또한 있었다[31-34]. 하지만, 단안깊이예측을 위한 데이터 증식 방법은 활발히 연구되지 않았으며 단지 몇몇 연구들 [31-34]에서만 행해진 것이 사실이다. [31] 연구의 CutDepth에서는 CutOut, CutMix와 같은 기존 2D Image Classification 연구에서 주로 활용되던 데이터증식 기법을 단안깊이예측에 활용하여 정확도를 올리고자 하였다. CutOut을 사용했을 경우 성능향상이 일어나지 않고, CutMix를 사용했을 경우 약 1.5%의 Abs. Rel 성능 향상을 이룰 수 있다고 분석하였다. [32-34] 연구에서는 2D Image Classification의 고전적 방법인 noise, brightness, contrast를 이용하여 단안깊이예측방식에 적용하여 정확도를 올리고자 하였다. 이에 대한 정확도는 KITTI 데이터셋 기준으로 0.112의 Abs. Rel 성능을 달성할 수 있다고 분석하였다.

본 연구에서는 단안깊이예측의 정확도를 높이기 위해 적은 수의 원본 데이터만으로도 데이터 증식을 통해 정확도를 높이는 방법에 대해 분석하였다. 특히 기존에 연구되어 졌던 방식들 scale, rotation, translation, noise, brightness control 등 에 더해 데이터 합성 방식들을 이용하여 이런 방식들이 단안깊이예측의 정확도 향상에 미치는 영향을 기술 별 기여도에 대해 분석하였

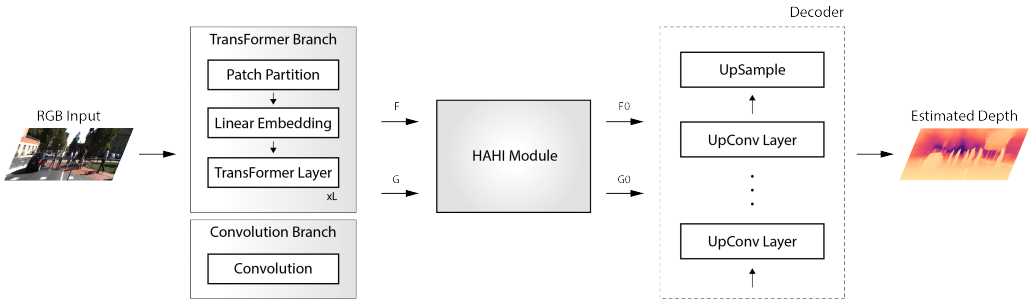


그림 1. DepthFormer의 구조도.
Fig. 1. Overall Architecture of DepthFormer.

다. 또한 손실함수 별로 단안깊이예측 정확도 향상에 최적화를 이룰 수 있는 방법을 실험적으로 분석하였다. 이는 지도학습 뿐 아니라 비지도학습, 준지도학습 모두에 적용 가능한 방법으로 로봇 및 자율주행 환경에서 현실적으로 단안깊이예측의 정확도를 올릴 수 있는 방법이다. 본 논문의 기여 포인트를 다음과 같이 정리하였다.

1) MDE성능향상을 위한 기존 데이터증식 기술분석:
기존 이미지 분류 작업에서 사용하던 고전적 데이터 증식 방식들을 단안깊이예측에 적용하여 성능 분석 및 성능 향상을 위한 기술 조합 도출

2) MDE성능향상을 위한 손실함수 분석, 제안:
단안깊이예측에서 사용되는 대표적 손실함수들을 적용하여 데이터 증식 기법 별 최적 손실함수 도출

3) MDE성능향상을 위한 신규 데이터증식 방식제안:
단안깊이예측을 위한 합성기반 방식인 Mask, Mask-Scale 방식을 제안하고 적용하여 성능 비교 분석

Ⅲ. 시스템 모델

실험으로 단안깊이예측의 다양한 데이터 증식 작업들에 대한 개별 성능 및 통합 성능을 확인하고 이를 기존 단안깊이예측 방식인 DepthFormer에 적용하여 성능을 확인하는 방향으로 진행되었다[35]. 우선 제안하는 단안깊이예측 방식 DepthFormer의 구조는 그림 1과 같다.

DepthFormer는 기본적으로 인코더-디코더 구조를 따른다. 이 구조는 입력 이미지에서 깊이 정보를 추정하기 위해 널리 사용되며, 입력 이미지를 저차원 특징으로 압축하고 다시 고해상도로 확장하는 역할을 한다. 인코더는 주로 EfficientNet, ResNet 및 DenseNet과 같은 특징 추출기를 활용하여 입력 이미지의 핵심적인 시각적 특징을 학습한다. 디코더는 convolution 및 upsampling 연산으로 구성되어 인코더에서 얻은 특징을 융합하여 해상도를 복원하며 고해상도 깊이맵을 예측한다. 또한 DepthFormer는 인코더와 디코더 사이에 HAHI(hierarchical aggregation and heterogeneous interaction) module을 제안하여 모델의 성능을 향상시킨다. HAHI module은 트랜스포머 Branch와 Convolution Branch로부터 얻은 특징 F, G간의 상호작용 및 관계를 모델링한다. 특히, 입력 데이터의 다양한 측면과 특징 간의 상호 작용을 강조함으로써 깊이 추정 작업에서 모델의 정확성을 향상시킨다.

Ⅳ. 데이터 증식

본 연구에서는 KITTI Dataset을 원본 데이터로 하여 제안하는 데이터 증식 기법들을 적용하여 실험하였다[36]. KITTI Dataset은 자율주행에 사용되는 데이터 셋으로 자동차, 보행자, 자전거 타는 사람 등의 클래스를 가지고 있다. 또한 깊이예측을 위해 필요한 깊이맵 데이터셋을 가지고 있다.

본 연구에서는 KITTI Dataset에 고전적 방식, 합성기반 방식의 데이터 증식 기법을 사용하여 정확도에 얼마만큼의 영향을 끼치는지 실험하였다. 데이터 증식은 다음의 방식에 기반하였다.

고전적 방식: Flip, Scale, Noise, Brightness
합성기반 방식: Mask, Mask-Scale

- Original : KITTI dataset을 사용하여 훈련을 진행
- Flip : 원본 데이터의 좌우를 반전시켜 데이터를 생성
- Scale : 원본 데이터의 중앙에서 가로, 세로로 각 각 1/2배 축소된 범위에서 crop하고 원본 이미지의 크기만큼 resize 하여 데이터를 생성
- Noise : 원본 데이터에 Gaussian Noise를 사용하여 데이터를 생성

(a) Flip (b) Scale

(c) Noise (d) Brightness

(e) Mask (f) Mask-Scale

그림 2. 데이터 증강을 적용한 KITTI 데이터 셋 예시
Fig. 2. Examples of KITTI Dataset with Data Augmentation.

- Brightness : 원본 데이터에서 밝기를 0.5배로 조정하여 데이터를 생성
- Mask : Segmentation 모델인 InternImage를 CityScape dataset으로 pretrained하고 객체를 Segmentation하여 Mask data를 생성[37,38]. 생성된 Mask data를 통해 마스크 연산을 진행하여 객체를 다른 사진에 붙이는 방법으로 데이터를 생성
- Mask-Scale : Mask 방식과 마찬가지의 방식으로 Segmentation을 통해 생성된 Mask data를 사용하였다. 추가적으로 Mask data와 원본 데이터의 가로, 세로를 각 각 1/2씩 축소시켜 마스크 연산을 수행하여 데이터를 생성

위에서 언급된 고전적 데이터 증식 방식인 Flip, Scale, Noise, Brightness 방식은 실제 주행환경에서 겪을 수 있는 다양한 환경 변화를 훈련 과정 중에 인위적으로 만들어줌으로서 실제 테스트 데이터 셋에서 더 높은 정확도 성능을 가능하게 하며 일반화 능력을 강화시키게 한다. 하지만, 객체 정보 자체에 대한 데이터 변이성은 부족하여 객체에 기반한 깊이예측 성능에는 한계를 갖는다.

반면 본 논문에서 새롭게 제안하는 합성 기반의 Mask 방식은 다른 이미지의 객체를 세그멘테이션 기술로 가져와 본 이미지에 합성하고 깊이 정보 또한 합성함으로서 객체에 기반한 깊이예측 데이터의 부족함을 보충해주어 성능향상에 더 기여할 수 있게 한다. 더 나아가 Mask-Scale 데이터 증식 방식은 Mask 방식의 한계점인, 기존 이미지의 객체 크기와 해당하는 객체 깊이 정보에 한정되어 증식이 되는 것에서 객체 크기와 해당 객체 깊이 정보를 비례적으로 변이 시켜 데이터를 증식시킴으로서 객체 자체 뿐 아니라 객체에 대한 깊이 정보 또한 합성으로 증식 시킬 수 있게 한다.

Ⅴ. 손실함수

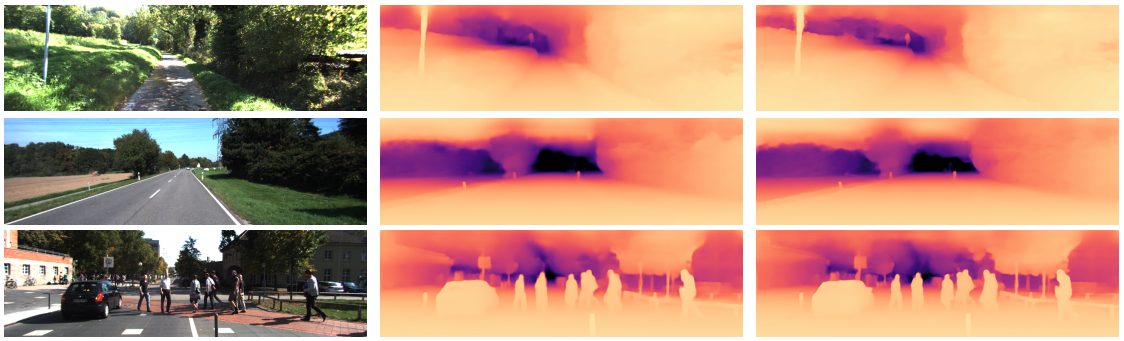
단안깊이예측을 위한 손실함수 들은 정답 깊이 값과 예측 깊이 값 간의 차이를 정량화 하여 손실 함수로 설정한다. 이를 위해 SigLoss, BerhuLoss를 손실함수로 사용하였다. 수식에서 사용된 d는 예측 값, d^*은 정답 깊이 값, c는 임계값을 의미하며 이들 수식은 다음과 같다.

5.1 SigLoss

$$d_i = \log(d) - \log(d^*) \tag{1}$$

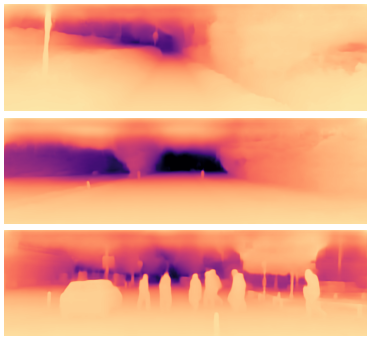
$$L_{SigLoss} = \frac{1}{T}\sum_i d_i^2 - \frac{1}{T^2}(\sum_i d_i)^2 \tag{2}$$

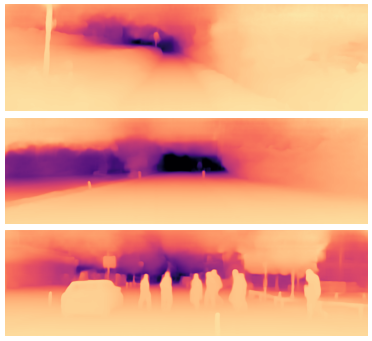
SigLoss(Scale Invariant Gradient Loss)는 깊이 예

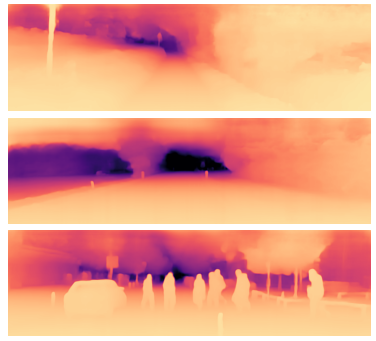

RGB

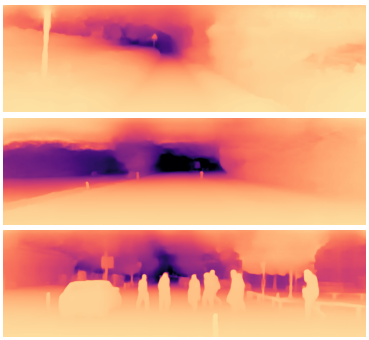
Original

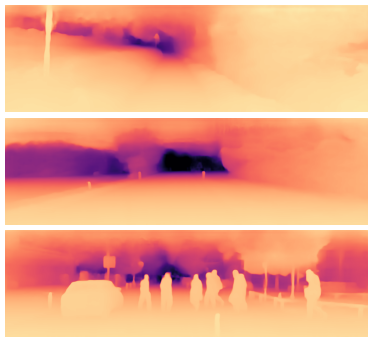
Original + Flip

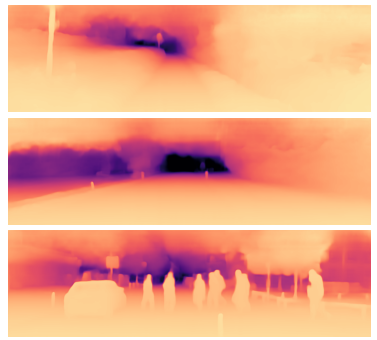

Original + Scale

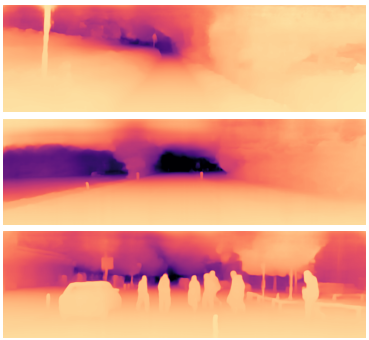

Original + Noise

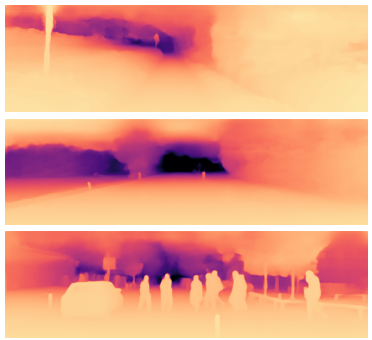

Original + bright


Original + Mask


Original + Mask-Scale


Original + Noise + Brightness


Original + Scale + Brightness


Original + Flip + Noise

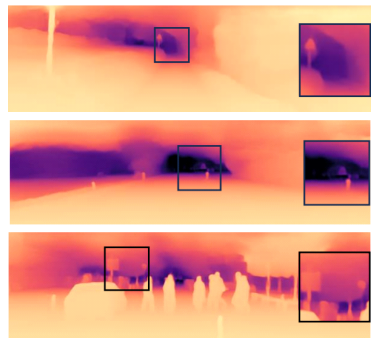

Original + Flip + Mask

그림 3. SigLoss를 사용한 Estimation 결과 예시
Fig. 3. Result Examples of Depth Estimation using SigLoss.

측의 절대적인 값에 의한 의존성을 줄이기 위해 설계되었다. 대부분의 깊이 추정 방법은 절대 깊이 값에 크게 의존하는데, 이러한 절대적인 깊이는 종종 잘못 추정될 수 있다. SigLoss는 이미지 내의 깊이 값 간의 상대적인 관계나 그래디언트에 초점을 맞추어 예측된 깊이 맵의 전반적인 구조가 원래의 깊이 맵과 유사하게 유지될 수 있다.

5.2 BerhuLoss

$$L_{Berhu}(d,d^*)=\begin{cases}|d-d^*| & \text{if } |d-d^*|\le c,\\ \dfrac{|d-d^*|^2+c^2}{2c} & \text{if } |d-d^*|>c.\end{cases} \quad (3)$$

Berhu loss는 임계값 c보다 작거나 같을 때는 $|d-d^*|$를 그대로 사용하지만 임계값 c보다 클 때, 즉 예측 오차가 임계값을 초과할 때는 상수 c^2을 더하고 $2c$로 나누어 이상치를 강하게 제거하면서도 일부 오차를 허용한다. 이를 통해 대량의 오차에 민감하지 않고, 모델을 안정적으로 학습시킨다.

Ⅵ. 실 험

실험을 위한 기기 환경으로는 2 way RTX 3090 기반의 NVIDIA GPU에 Pytorch 기반으로 코딩하여 실험결과를 확인하였다. 학습률 조정 방식으로는 CosineAnnealing을 사용하였으며, Original data로 72084 쌍의 RGB image와 Depthmap을 사용하고 추가적인 Augmentation 기법 당 72084 쌍을 추가로 사용하여 훈련을 진행하였다. 2 batch size로 38,400 iteration 동안 훈련을 진행하여 최적의 성능을 도출하였다. 테스트 이미지의 추론 지연 시간은 각 이미지 당 0.222초의 시간이 소요되어 초당 약 4.5 프레임을 기록하였다.

성능 파라미터로서 사용한 REL은 Absolute Relative Error로 수식 (1)에서 보듯이 특정 픽셀 p에 대하여 예측한 값 $\hat{d}_p$와 실제 값 d_p의 차이를 실제 값 d_p으로 나눈 것을 모두 더하여 d_p, $\hat{d}_p$이 모두 존재하는 픽셀의 총 개수 T로 나눈 지표로써, 예측한 픽셀에서 Depth를 얼마나 재현하였는지를 나타낸다.

$$\frac{1}{T}\sum_{p}\frac{|d_p-\hat{d}_p|}{d_p} \quad (4)$$

이 연구에서 우리는 여러 가지 데이터 증강과 조합을 통해 단안깊이예측 모델인 DepthFormer의 성능을 분석하고 증대시킬 수 있는 손실함수와 Data Augmentation 방식에 대해 확인하였다.

먼저 SigLoss를 손실함수로 하여 원본 데이터 만을 사용하여 훈련한 결과, 0.0528의 REL을 도출하였다. 단일 Augmentation Data을 Original Data와 합쳐 훈련한 결과, 표 1과 같은 결과를 얻을 수 있었다. 단일 Augmentation Data를 포함하여 훈련할 경우 대부분의 증강 기법에서 유의미한 성능 향상을 보이지는 않았으며, 단일 Augmentation에서는 Original + Mask-scale이 0.0513의 REL으로 최적의 성능을 나타냈다. 복합 Augmentation에서도 마찬가지로 Noise와 Brightness, Scale과 Brightness, Flip과 Noise를 조합하였을 때, 각각 0.0559, 0.0561, 0.0533 REL로 오히려 성능이 저하되는 것을 보였다. 하지만 Flip과 Mask를 조합하였을 때, 최적의 결과인 0.0508의 REL을 도출하였다. 이는 원래의 KITTI dataset만으로 훈련시켰을 때의 결과인

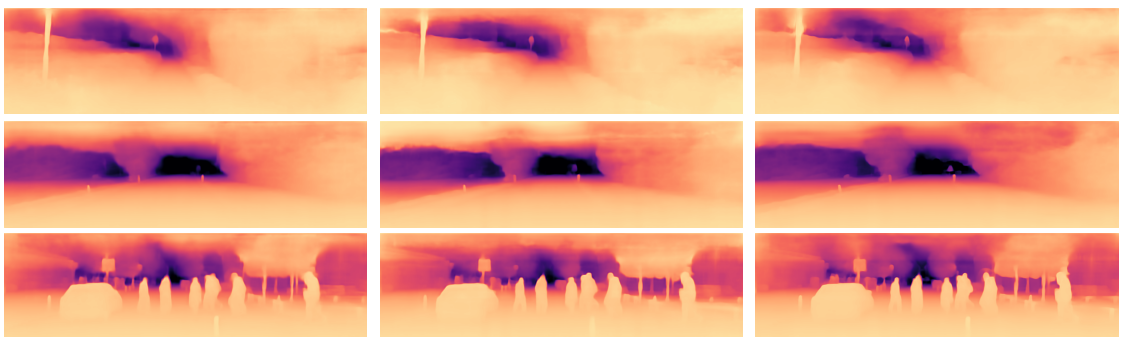


그림 4. BerhuLoss를 사용한 Depth Estimation 결과 예시
Fig. 4. Result Examples of Depth Estimation using BerhuLoss.

표 1. 데이터 증강 별 성능 결과
Table. 1. Performance Results by Data Augmentation.

Loss Function	Augmentation	REL ↓
SigLoss	Original	0.0528
	Original + Flip	0.0534
	Original + Scale	0.057
	Original + Noise	0.0551
	Original + Brightness	0.0562
	Original + Mask	0.0538
	Original + Mask-Scale	0.0513
	Original + Noise + Brightness	0.0559
	Original + Scale + Brightness	0.0561
	Original + Flip + Noise	0.0553
	Original + Flip + Mask	0.0508
BerhuLoss	Original	0.0595
	Original + Flip	0.0599
	Original + Scale	0.0643
	Original + Noise	0.0619
	Original + Brightness	0.0638
	Original + Mask	0.0612
	Original + Mask-Scale	0.0586
	Original + Noise + Brightness	0.0628
	Original + Scale + Brightness	0.0632
	Original + Flip + Noise	0.0619
	Original + Flip + Mask	0.0581

0.0528 REL보다 3.9% 만큼 향상된 성능이다. 그림 3는 각 각의 원본과 증강된 데이터를 통해 예측한 depth 이미지이다. 원본으로 예측한 depth에서는 표현하지 못했던 멀리 있는 표지판의 존재나 형태를 Original + Flip + Mask 데이터 셋을 통해 예측한 depth에서는 명확하게 표현하는 것을 볼 수 있다. 제안된 BerhuLoss를 손실함수로 사용하여 훈련시킨 결과 SigLoss를 사용했을 때보다 성능은 저하되었지만 사용한 Augmentation 기법에 따라 REL이 유사한 비율로 나오는 것을 알 수 있었다. 그림 4은 BerhuLoss를 사용하여 예측한 깊이 맵이다. 비교를 위해 원본, 가장 좋지 않은 케이스, 가장 좋은 케이스인 Original, Original + Scale, Original + Flip + Mask 데이터로 훈련된 가중치를 사용하였다.

Ⅶ. 결 론

본 연구에서는 단일 카메라로 주변 환경의 깊이 정보를 얻어내는 단일깊이예측 방식의 성능을 최적화하는 방법에 대해 연구하였다. 실험을 통해 고전적 방식과 합성기반 방식의 Augmentation을 적절히 조합하여 훈련하였을 때 깊이예측의 성능 향상에 도움이 되는 것을 알 수 있었다. 특히 제안된 Mask 기반 Augmentation을 포함한 데이터 셋을 사용하여 학습할 때, 동일한 데이터 크기에도 불구하고 우수한 성능을 달성할 수 있었다. 하지만 개별적인 데이터 증강 기법을 독립적으로 적용하는 것은 성능이 저하시킬 수 있음을 확인하였다. 또한 실험에 사용한 두개의 손실함수 중, SigLoss가 전반적으로 더 나은 REL값을 나타내었다. 구체적으로, SigLoss를 사용할 때 모델의 깊이 추정 정확도가 향상되었으며, 이미지가 다양한 영역에서 세부적인 깊이 패턴을 더 잘 포착할 수 있었다. 반면 BerhuLoss를 사용한 경우에는 비교적 높은 REL값이 관찰되었으나, 전체적인 깊이 구조는 잘 유지되었다. 이러한 결과는 SigLoss가 BerhuLoss에 비해 깊이 추정 작업에서 더 우수한 성능을 나타낼 수 있음을 시사한다.

이러한 결과들은 자율주행 차량 및 로봇 등 실제 응용 환경에서도 유용하게 적용될 수 있을 것으로 기대된다. 특히 데이터 레이블링 비용이 높고 추가 데이터 확보가 어려운 상황에서 적은 비용으로 깊이예측 모델의 성능을 향상시키는데 큰 도움을 줄 것으로 기대된다.

References

[1] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *J. Field Robotics*, vol. 37, no. 3, pp. 362-386, 2020. (https://doi.org/10.1002/rob.21918)

[2] A. I. Károly, P. Galambos, J. Kuti, and I. J. Rudas, “Deep learning in robotics: Survey on model structures and training strategies,” *IEEE Trans. Syst., Man, and Cybernetics: Syst.*, vol. 51, no. 1, pp. 266-279, 2020. (https://doi.org/10.1109/TSMC.2020.3018325)

[3] A. Bhoi, *Monocular depth estimation: A survey*, Jan., 27, 2019, from *arXiv: 1901.09402*, 2019. (https://doi.org/10.48550/arXiv.1901.09402)

[4] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, “Monocular depth estimation using deep learning: A review,” *Sensors*, vol. 22, no. 14, 5353, 2022. (https://doi.org/10.3390/s22145353)

[5] R. Xiaogang, Y. Wenjing, H. Jing, G. Peiyuan, and G. Wei, "Monocular depth estimation based on deep learning: A survey," in *IEEE 2020 Chin. Automat. Congress (CAC),* pp. 2436-2440, Nov. 2020. (https://doi.org/10.1109/CAC51589.2020.9327548)

[6] Y. M. Tsai, Y. L. Chang, and L. G. Chen, "Block-based vanishing line and vanishing point detection for 3d scene reconstruction," in *IEEE 2006 Int. Symp. Intell. Sign. Process. and Commun.,* pp. 586-589, 2006. (https://doi.org/10.1109/ISPACS.2006.364726)

[7] C. Tang, C. Hou, and Z. Song, "Depth recovery and refinement from a single image using defocus cues," *J. Modern Optics*, vol. 62, pp. 441-448, 2015. (https://doi.org/10.1080/09500340.2014.967321)

[8] R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah, "Shape-from shading: A survey," *IEEE Trans. Pattern Analysis and Mach. Intell.*, vol. 21, pp. 690-706, 1999. (https://doi.org/10.1109/34.784284)

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in NIPS*, vol. 25, 2012. (https://doi.org/10.1145/3065386)

[10] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, Sep., 4, 2014, from *arXiv: 1409.1556.* (https://doi.org/10.48550/arXiv.1409.1556)

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR,* pp. 770-778, 2016. (https://doi.org/10.1109/cvpr.2016.90)

[12] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in NIPS*, pp. 2366-2374, 2014. (https://doi.org/10.48550/arXiv.1406.2283)

[13] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: Camera-aware multi-scale convolutions for single-view depth," in *Proc. IEEE Conf. CVPR*, pp. 11826-11835, 2020. (https://doi.org/10.1109/cvpr.2019.01210)

[14] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. CVPR*, pp. 270-279, 2017. (https://doi.org/10.1109/cvpr.2017.699)

[15] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *IEEE 2016 Fourth Int. Conf. 3D Vision (3DV),* pp. 239-248, 2016. (https://doi.org/10.1109/3dv.2016.32)

[16] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. CVPR*, pp. 5162-5170, 2015. (https://doi.org/10.1109/cvpr.2015.7299152)

[17] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. CVPR*, pp. 1983-1992, 2018. (https://doi.org/10.1109/cvpr.2018.00212)

[18] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE Conf. CVPR*, pp. 340-349, 2018. (https://doi.org/10.1109/cvpr.2018.00043)

[19] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proc. IEEE Conf. CVPR*, pp. 9788-9798, 2019. (https://doi.org/10.1109/cvpr.2019.01002)

[20] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE Conf. CVPR*, pp. 3917-3925, 2018. (https://doi.org/10.1109/cvpr.2018.00412)

[21] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal

estimation from monocular images using regression on deep features and hierarchical crfs," in *Proc. IEEE Conf. CVPR*, pp. 1119-1127, 2015. (https://doi.org/10.1109/cvpr.2015.7298715)

[22] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. CVPR*, pp. 2800-2809, 2015. (https://doi.org/10.1109/cvpr.2015.7298897)

[23] E. Ricci, W. Ouyang, X. Wang, N. Sebe, et al., "Monocular depth estimation using multi-scale continuous crfs as sequential deep networks," *IEEE Trans. Pattern Analysis and Mach. Intell.*, vol. 41, pp. 1426-1440, 2018. (https://doi.org/10.1109/tpami.2018.2839602)

[24] A. Mousavian, H. Pirsiavash, and J. Košecká, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *IEEE 2016 Fourth Int. Conf. 3D Vision (3DV),* pp. 611-619, 2016. (https://doi.org/10.1109/3dv.2016.69)

[25] J. Sun, N. N. Zheng, and H. Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Analysis and Mach. Intell.*, vol. 25, pp. 787-800, 2003. (https://doi.org/10.1109/tpami.2003.1206509)

[26] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Comput. Soc. Conf. CVPR 2004*, pp. 964-971, 2004. (https://doi.org/10.1109/cvpr.2004.1315094)

[27] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," pp. 1-16, 2020. (https://doi.org/10.1007/978-3-030-58529-7_34)

[28] X. Ye, X. Ji, B. Sun, S. Chen, Z. Wang, and H. Li, "Drm-slam: Towards dense reconstruction of monocular slam with scene depth fusion," *Neurocomputing*, vol. 396, pp. 76-91, 2020. (https://doi.org/10.1016/j.neucom.2020.02.044)

[29] W. Zhao, S. Zhang, Z. Guan, H. Luo, L. Tang, J. Peng, and J. Fan, "6D object pose estimation via viewpoint relation reasoning," *Neurocomputing*, vol. 389, pp. 9-17, 2020. (https://doi.org/10.1016/j.neucom.2019.12.108)

[30] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. CVPR*, pp. 1851-1858, 2017. (https://doi.org/10.1109/cvpr.2017.700)

[31] Y. Ishii, T. Yamashita, *CutDepth: Edge-aware Data Augmentation in Depth Estimation*, Jul., 16, 2021, from arXiv:2107.07684. (https://doi.org/10.48550/arXiv.2107.07684)

[32] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 3828-3838, 2019. (https://doi.org/10.1109/iccv.2019.00393)

[33] S. Pillai, R. Ambrus, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 ICRA IEEE*, pp. 9250-9256, 2019. (https://doi.org/10.1109/icra.2019.8793621)

[34] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proc. IEEE/CVF Conf. CVPR*, pp. 4756-4765, 2020. (https://doi.org/10.1109/cvpr42600.2020.00481)

[35] Z. Li, Z. Chen, X. Liu, and J. Jiang, *Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation*, Mar., 27, 2022, from *arXiv:2203.14211.* (https://doi.org/10.1007/s11633-023-1458-0)

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conf. CVPR*, pp. 3354-3361, Jun. 2012. (https://doi.org/10.1109/cvpr.2012.6248074)

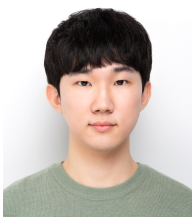
[37] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, and Y. Qiao, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. CVPR*, pp. 14408-14419, 2023.

(https://doi.org/10.1109/cvpr52729.2023.01385)
[38] N. Gählert, N. Jourdan, M. Cordts, U. Franke, and J. Denzler, *Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection*, Jun., 14, 2020, from arXiv:2006.07864. (https://doi.org/10.48550/arXiv.2006.07864)

전 우 민 (Woomin Jun)

2021년 3월~현재 : 동서울대학교 전자공학과 전문학사과정
<관심분야> 딥러닝, 영상인식

유 지 상 (Jisang Yoo)



2023년 2월 : 동서울대학교 전자공학과 졸업
2023년 3월~현재 : 동서울대학교 전자공학과 학사과정
<관심분야> 딥러닝, 영상인식

이 성 진 (Sungjin Lee)

2011년 8월 : 연세대학교 전자공학과 박사 졸업
2012년 9월~2016년 7월 : 삼성전자 DMC연구소 책임연구원
2016년 7월~현재 : 동서울대학교 전자공학과 조교수
<관심분야> 딥러닝, 영상인식, 3D Reconstruction,
[ORCID:0000-0003-3159-8394]