JPNT 10(4), 297-305 (2021)
https://doi.org/10.11003/JPNT.2021.10.4.297

JPNT Journal of Positioning, Navigation, and Timing

Deep Learning Based Monocular Depth Estimation: Survey

Chungkeun Lee, Dongseok Shim, H. Jin Kim†

Department of Aerospace Engineering, Seoul National University, Seoul 08826, Korea

ABSTRACT

Monocular depth estimation helps the robot to understand the surrounding environments in 3D. Especially, deep-learning-based monocular depth estimation has been widely researched, because it may overcome the scale ambiguity problem, which is a main issue in classical methods. Those learning based methods can be mainly divided into three parts: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning trains the network from dense ground-truth depth information, unsupervised one trains it from images sequences and semi-supervised one trains it from stereo images and sparse ground-truth depth. We describe the basics of each method, and then explain the recent research efforts to enhance the depth estimation performance.

Keywords: deep learning, depth estimation

1. 서론

로보틱스 분야에서 카메라는 주변 환경을 파악하는데 있어서 가볍고 저렴한 센서로 많은 각광을 받고 있다. 카메라로부터 측정된 이미지는 주변 환경에 대한 많은 정보를 가지고 있다는 장점이 있으나, 측정된 이미지로부터 의미 있는 정보를 추출하기 위해 추가적인 노력이 필요하다는 단점이 있다. 이러한 연유로 이미지로부터 의미 있는 정보인 카메라의 3차원 위치, 측정된 이미지 픽셀의 3차원 정보 및 깊이 등을 추정하는 연구가 많이 진행되어 왔다.

특히, 로봇은 3차원 환경에서 동작하므로, 이미지 정보를 3차원 정보로 변환하는 노력이 많이 이루어지고 있다. 대표적으로는 이미지로부터 3차원 위치를 추정하는 기법으로 카메라 자신의 위치를 추정하는 오도메트리 (odometry) 기법, 카메라로부터 측정된 이미지 픽셀 각각의 3차원 위치를 추정하는 깊이 추정 (depth estimation) 기법, 그리고 자신의 위치와 더불어 주변 환경을 파악하는 Simultaneously Localization and Mapping (SLAM) 기법이 있다.

Received Aug 30, 2021 **Revised** Nov 08, 2021 **Accepted** Nov 17, 2021

†**Corresponding Author**
E-mail: hjinkim@snu.ac.kr
Tel: +82-2-880-1552 Fax: +82-2-888-0321

Chungkeun Lee https://orcid.org/0000-0003-1232-0901
Dongseok Shim https://orcid.org/0000-0003-1580-0716
H. Jin Kim https://orcid.org/0000-0002-6819-1136

하지만, 카메라는 구조적인 한계에 의하여 스케일 모호성을 가지며, 이는 특히 단안 카메라만을 이용하여 3차원 공간 정보를 복원하는 것을 어렵게 한다. 이러한 스케일 모호성을 해결하기 위해 RGB-D, IMU, 스테레오 카메라 등의 추가 센서를 활용하여 물리적으로 모호성을 해결하는 연구가 진행되어 왔다.

이러한 추가 센서는 비용 및 무게 적인 측면에서 손해이며, 센서 사이의 동기화 및 외부 파라미터를 추정하는 등의 추가적인 과정이 필요하다. 이러한 이유로 인해, 단안 카메라만을 이용한 방법 역시 꾸준히 제안되어 왔다. 다만, 이러한 방법은 연속적인 이미지가 주어진 환경에서, 실제 스케일 정보를 추정하는 것을 포기하고 픽셀 사이의 깊이 비율을 추정하고, 프레임 사이의 스케일을 유지하여 프레임 사이의 일관성을 가지도록 하는 선에서 만족하게 된다 (Forster et al. 2014, Mur-Artal et al. 2015, Ranftl et al. 2016).

학습 기반 깊이 추정 기법은 단안 카메라의 한계로 지적되는 스케일 문제를 해결하는 깊이 추정 방식으로서 각광받고 있다. 사전에 수집된 깊이 참값 정보를 바탕으로 스케일 정보가 포함된 깊이 정보를 컨볼루션 신경망 (convolutional neural network)을 이용하여 학습하는 기법이다. 이러한 기법은 깊이 스케일을 제공하고, 깊이 정보 역시 전체 픽셀에 대해 제공한다는 장점이 있다.

본 논문은 학습 기반 단안 카메라 깊이 추정의 연구들을 학습 데이터에 따라 분류하고, 해당 방법에 대한 대략적인 설명 및 연구 동향에 대하여 제시한다. 이와 함께, 학습에 일반적으로 활용

하는 데이터 셋 및 검증 지표를 설명하고, 각 방법의 정확도를 정리하여 제시하여 깊이 추정 기법들의 정확도 비교와 더불어 어떠한 방법으로 정확도를 높였는지를 요약하는데 목적을 둔다.

2. 학습 기반 단안 카메라의 깊이 추정 기법

학습 기반 단안 카메라의 깊이 추정 기법은 학습 데이터로부터 신경망을 학습하며, 추후 구동시에는 단일 이미지와 학습된 신경망을 활용하여, 전체 픽셀 정보에 대해 깊이 값이 주어지는 조밀한 깊이 정보 (dense depth map)를 추정하는 방식이다. 학습 데이터로 어떤 데이터를 활용하는지, 그리고 신경망 구조를 어떻게 구성하고, 어떤 방식의 손실 함수 (loss function)을 설계하는지에 따라 여러가지 기법으로 나뉜다.

단안 카메라의 깊이 추정은 공통적으로 구동 (inference) 단계에서는 단안 이미지만을 활용하여 깊이 정보를 추정하게 된다. 그러한 연유로 학습하는 신경망은 활용하는 학습 데이터와 무관하게, 단안 이미지만을 입력으로 받으며 해당 이미지의 깊이 정보를 출력으로 한다.

가장 기본적인 기법은 지도식 기법 (supervised method)으로, 학습 정보로 이미지 및 조밀한 깊이 참값이 쌍으로 주어지는 환경에서, 깊이 참값을 활용하여 신경망의 학습을 수행하는 방법이다. 이는 깊이 참값을 추정하는 네트워크를 보간 (regression)하는 문제로, 어떻게 네트워크를 구성하고 어떠한 방식으로 보간 하는가에 따라 다양한 방법이 제안되었다 (Eigen et al. 2014, Laina et al. 2016, Fu et al. 2018).

지도식 기법은 깊이 참값 정보를 수집해야 하므로, 깊이 참값 정보를 수집하기 위해 추가적인 센서를 활용하여야 한다. 일반적으로 실내 환경에서는 일반적으로 RGB-D 카메라를, 실외 환경에서는 라이다 (LIDAR)를 많이 활용한다. 라이다의 경우 조밀하지 않은 (sparse) 깊이 참값이 수집되며, 이를 보간 하여 조밀한 깊이 참값 정보를 생성하게 된다.

하지만 깊이 참값을 수집하는데 추가적인 노력이 필요하며, 특히 야외 환경에서는 값비싼 라이다 센서를 활용하여야 한다. 이러한 어려움을 해결하기 위해, 깊이 참값 정보 없이 학습하는 비지도식 기법 (unsupervised method)이 제안되었다. 초창기의 비지도식 기법으로는 스테레오 이미지를 학습 데이터로 하며, 스테레오 카메라의 에피폴라 기하학 (epipolar geometry)을 기반으로 학습을 수행하는 방법이다. 이러한 방법은 깊이 참값을 수집하지 않아도 된다는 장점이 있다 (Garg et al. 2016, Godard et al. 2017).

이러한 연구는 단안 카메라의 연속적인 이미지를 기반으로 움직이는 카메라에 대해 에피폴라 기하학을 적용시키는 방향으로 진행되었다 (Zhou et al. 2017, Yin & Shi 2018, Godard et al. 2019). 이 방법은 학습 데이터로 단안 카메라의 연속적인 이미지만을 활용하기에, 학습 데이터와 검증 데이터의 차이가 거의 없다는 장점이 있다. 다만, 이 방법은 스케일 모호성을 해결하는 정보가 제공되지 않아서, 기존 방법과 마찬가지로 스케일 모호성을 해결하지 못한다는 문제점이 있다.

최근 연구에서는 이러한 두가지 비지도식 기법을 구분하기 위하여, 스테레오 카메라를 학습 데이터로 활용하는 기법을 준지도식 기법 (semi-supervised learning) 혹은 자기지도식 기법 (self-supervised learning)으로 구분하는 경우가 많다. 본 논문에서도 마찬가지로 스테레오 카메라를 학습 데이터로 활용하는 경우를 준지도식 기법으로 구분한다. 이와 함께, 조밀하지 않은 깊이 정보를 활용하는 등 조밀한 깊이 참값이 제공되지 않으나 단안 이미지 이외의 다른 학습 데이터가 필요한 경우 역시 준지도식 기법으로 구분한다.

위 분류와 무관하게, 모든 기법에 활용이 가능한 네트워크 구조 및 학습 방법에 대한 연구 역시 존재한다. 가상의 이미지 및 깊이 정보는 생성하기 쉽다는 점을 착안하여, 가상의 데이터셋을 통하여 학습하고 이를 실제 환경으로 전이 (transfer)하여 학습하는 도메인 적응 기법 (domain adaptation) (Kundu et al. 2018, Zheng et al. 2018, Zhao et al. 2019), 혹은 생성적 적대 신경망 (generative adversarial network; GAN)을 활용하는 경우 (Aleotti et al. 2018, Almalioglu et al. 2019), 또는 신경망 구조를 개선하여 순환 신경망 (recurrent neural network)을 활용하는 등의 방법론 (Wang et al. 2019) 역시 제안되었다.

2.1 지도식 깊이 추정 기법

지도식 깊이 추정 기법은 깊이 참값 정보가 포함된 학습 데이터로부터 신경망을 학습하며, 추후 구동시에는 단일 이미지와 학습된 신경망을 활용하여 조밀한 깊이 정보를 추정하는 방식이다. 학습 데이터로 어떤 데이터를 활용하는지, 그리고 신경망 구조를 어떻게 구성하고, 어떤 방식의 손실 함수를 설계하는지에 따라 여러가지 기법으로 나뉜다.

지도식 학습 기법은 깊이 참값 정보를 활용한다. 그러므로, 깊이 참값 d_i과 신경망을 통해 추정된 깊이 $\hat{d}_i$의 거리를 최소화하도록 학습을 수행한다. 일반적으로 적절한 거리 함수 $\mathrm{dist}(\cdot,\cdot)$에 대하여, Eq. (1)과 같은 손실 함수를 정의하여, 손실 함수를 최소화한다.

$$L = \frac{1}{n}\sum_{i=1}^{n} \mathrm{dist}(\hat{d}_i, d_i) \tag{1}$$

Eigen et al. (2014)은 2단계의 컨볼루션 신경망을 구성하여, 지도식 학습을 수행하는 방법을 제안하였다. 신경망을 학습하기 위한 손실 함수로 스케일 $\lambda \in [0,1]$에 대하여 스케일 불변 오차를 정의하였다 (Eq. (2)). 해당 기법은 최초로 제안된 end-to-end 기법의 단안 깊이 추정 방식으로, 학습 기반의 깊이 추정이 가능함을 보였다는데 큰 의의가 있다.

$$L = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log \hat{y}_i)^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n}(\log y_i - \log \hat{y}_i)^2\right)^2 \tag{2}$$

Laina et al. (2016)은 깊이 추정을 위한 컨볼루션 신경망으로 기존의 컨볼루션 구조와 완전 연결 (fully connected) 구조를 병합한 형태에서, ResNet (He et al. 2016) 신경망에 기반한 활용한 완전 컨볼루션 (fully convolutional) 구조를 제안하였다.

Fu et al. (2018)은 깊이 추정 문제를 순서형 회귀분석 (ordinal

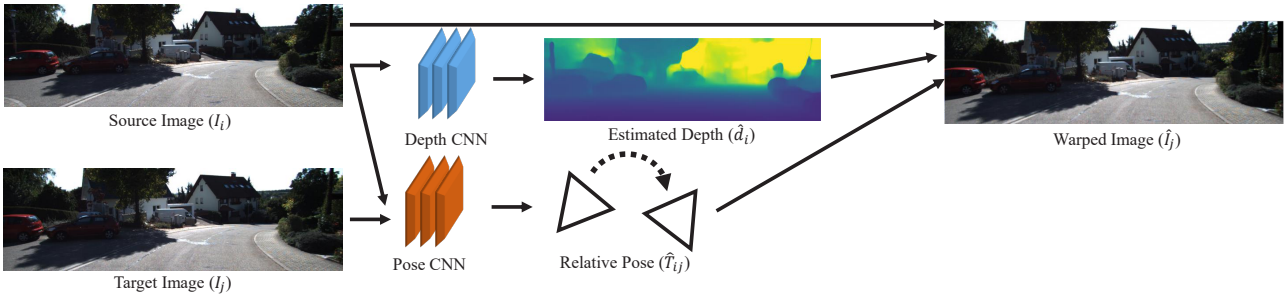

Fig. 1. The diagram of training step in unsupervised monocular depth estimation. From source image I_i, depth CNN estimates depth $\hat{d}_i$ in source view. From both source and target images (I_i, I_j), pose CNN estimates relative pose $\hat{T}_{ij}$. With estimated depth $\hat{d}_i$ and relative pose $\hat{T}_{ij}$, the source image is warped as in Eqs. (4) and (5) in target view denoted as $\hat{I}_j$. Then, the loss function is defined as the difference between the target image I_j and warped image $\hat{I}_j$ as in Eq. (6).

regression) 문제로 변환하고, 이에 알맞은 신경망 구조 및 손실 함수를 제안하였다. 해당 방법론은 제안 당시 최고의 깊이 추정 성능을 보였던 것으로 알려져 있다.

Wofk et al. (2019)은 신경망 네트워크를 구성하는데 있어서 MobileNet (Howard et al. 2017)에 기반하여 임베디드 환경에서 실시간으로 작동하는 깊이 추정 구조를 제안하였다. 이때, 구동 속도 증가에 비하여 깊이 추정 성능의 하락은 크지 않음을 보임으로 임베디드 환경에서 학습 기반 깊이 추정 기법의 적용 가능성을 보였다.

Lee & Kim (2019)은 조밀한 깊이 정보를 스케일 및 여러 요소의 스케일 불변한 상대 깊이 (relative depth)로 나누어 표현하고, 각각의 상대 깊이를 추정하는 신경망 구조를 제안하였다. 깊이 정보 $D_n \in \mathbb{R}^{2^n \times 2^n}$에 대하여, 상대 깊이 $F_n \in \mathbb{R}^{2^n \times 2^n}$는 주변 4개의 픽셀의 기하평균으로 나눈 값으로 Eq. (3)과 같이 정의된다.

$$F_n(2i,2j) = \frac{D_n(2i,2j)}{\prod_{k=0}^{1}\prod_{l=0}^{1} D_n(2i-k,2j-l)^{\frac{1}{4}}} \tag{3}$$

지도식 깊이 추정 기법은 Eigen et al. (2014)에서 가능성을 보인 것을 시작으로, 성능을 향상시키기 위한 연구가 주로 이루어졌다. 이에 대한 해결책으로, 신경망의 구조를 개선하거나 (Laina et al. 2016, Lee & Kim 2019), 학습 방법을 개선하는 (Fu et al. 2018) 연구가 많이 제안되었다. 다만 일부 연구는 딥러닝의 작동 시간에 초점을 맞추어, 로봇 등에 활용하기 위한 실시간 구현에 기여하는 경우도 있다 (Wofk et al. 2019).

2.2 비지도식 깊이 추정 기법

비지도식 깊이 추정 기법은 깊이의 참값 정보 없이 연속적으로 측정된 이미지만을 활용하여 깊이 추정 신경망을 학습하는 방식이다. 비지도식 깊이 추정 기법은 두 시점에서 촬영된 이미지의 카메라 자세 및 깊이 정보가 주어지는 경우, 에피폴라 기하학에 의거 3차원 좌표 변환을 통해 시점 변환이 가능하다는 점을 이용한다. 이러한 정보를 활용하기 위해, 학습 데이터는 연속된 이미지로 주어져야 한다. 또한, 학습 및 검증 과정에서 해당 이미지의 깊이 정보 d_i 뿐만 아니라, 두 이미지 사이의 상대적인 6-dof 자세 정보인 T_{ij} 역시 추정한다 (Fig. 1).

i 시점에서의 픽셀 좌표 $u_i \in \mathbb{R}^3$, 해당 좌표에서의 깊이 정보 $d_i \in \mathbb{R}$, 카메라 내부 파라미터 $K \in \mathbb{R}^{3\times3}$가 주어지는 경우, 3차원 좌표 $X_i \in \mathbb{R}^3$는 Eq. (4)와 같이 표현된다. $i \rightarrow j$의 자세 정보 $T_{ij} = (R_{ij}, p_{ij}) \in \mathbb{R}^{3\times3} \times \mathbb{R}^3$을 활용하면, j 시점에서의 3차원 좌표 $X_j \in \mathbb{R}^3$ 및 픽셀 좌표 $u_j \in \mathbb{R}^3$를 Eq. (5)와 같이 표현할 수 있다. 이러한 좌표 변환을 활용하여, 카메라 내부 파라미터, i 시점에서의 깊이 정보 및 $i \rightarrow j$의 자세 정보를 활용하면, i 시점에서의 이미지를 j 시점의 픽셀 좌표로 변환할 수 있다.

$$X_i = d_i K^{-1} u_i \tag{4}$$

$$u_j \sim K X_j = K(R_{ij} X_i + p_{ij}) \tag{5}$$

여기서 변환된 픽셀 좌표 u_j는 자연수에서 정의되는 일반적인 픽셀 좌표가 아니라 실수에서 정의된다. 이를 해결하기 위하여, 미분가능한 쌍일차 보간 (differentiable bilinear sampling) (Jaderberg et al. 2015)을 통하여 시점에서 측정된 이미지 I_i를 j 시점에서 측정되었을 이미지 $\hat{I}_j$로 변환한다. 이렇게 변환된 이미지와 실제 j 시점에서 측정된 이미지 I_j와의 거리 Eq. (6)이 최소화되도록 학습을 수행한다.

$$L = \frac{1}{n}\sum_{i=1}^{n} \mathrm{dist}(\hat{I}_j, I_j) \tag{6}$$

이러한 비지도식 학습 기법은 스케일 모호성 문제를 해결하지 못한다는 문제점이 있다. 예를 들어, 임의의 스케일 $\lambda \in (0,\infty)$에 대하여, 깊이 정보와 좌표 변환이 (d_i, T_{ij})가 아니라, Eq. (7)에서 주어진 깊이 및 자세 쌍(d_i', T_{ij}')으로 추정된 경우에도 j 시점의 픽셀 좌표 u_j는 동일한 값을 가진다.

$$d_i' = \lambda d_i, T_{ij}' = \left(R_{ij}, \frac{p_{ij}}{\lambda}\right) \tag{7}$$

Zhou et al. (2017)은 연속된 이미지에 대하여 깊이 정보와 자세 정보를 동시에 추정하여 학습하는 비지도식 학습 기법을 최초로 제안하였다. 이를 통하여 깊이 추정 신경망을 학습하는데 있어서 비지도식 기법으로 학습이 가능하다는 것을 보였다.

Yin & Shi (2018)는 깊이 정보 및 자세 정보에 더불어 광 흐름 (optical flow)을 동시에 추정하는 구조를 제안하였다. 이렇게 추

정한 광 흐름을 기반으로 깊이 정보의 일관성을 드러내는 손실 함수를 추가하여서, 깊이 추정의 정확도를 향상하였다.

Godard et al. (2019)은 기존의 가림 및 물체 이동을 고려하기 위해 마스크를 추정하는 대신, 두 이미지의 차이를 기반으로 한 자동 마스크 (auto-mask) 기법을 제안하여, 깊이 추정에 활용되는 네트워크의 구조를 단순화 시킴과 동시에 깊이 추정의 정확도를 향상하는 기법을 제시하였다.

Almalioglu et al. (2019)은 생성적 적대 신경망 기법을 비지도식 방식에 적용하여 깊이 추정 및 자세 추정 정확도를 향상하였다. 생성적 적대 신경망 기법은 가상의 이미지를 생성하는 생성 신경망 및 주어진 이미지가 실제 이미지인지 생성된 이미지인지를 학습하는 분류 신경망을 동시에 학습하는 기법이다. 해당 연구에서는 생성 신경망에서는 깊이 정보를 추정하여 이를 Eqs. (4)와 (5)를 이용하여 변환하고, 분류 신경망에서 변환된 이미지와 실제 이미지를 분류하는 방식을 제안하였다.

Wang et al. (2019)은 컨볼루션 순환 신경망을 이용한 깊이 추정 네트워크를 제안하였다. 제안된 네트워크는 연속적인 이미지를 활용하여 학습 및 구동을 수행하는 경우의 시계열 정보를 학습하기 용이하다는 장점이 있다. 이를 통하여 연속적인 이미지를 활용하는 환경에서 깊이 추정 정확도를 향상시켰다.

비지도식 학습 기법은 Zhou et al. (2017)에서 가능성이 제시된 것을 시작으로, 성능 향상을 위한 연구가 많이 수행되었다. 비지도식 학습 기법의 연구는 크게 세가지의 연구 방향이 존재한다. 첫째로, 일반적으로 카메라 자세 및 깊이 정보가 아닌 다른 정보 역시 추정하고 이를 활용하여 성능을 높이고자 하는 연구이다 (Yin & Shi. 2018). 둘째로, 개선된 학습 방법을 제안하여 성능을 높이고자 하는 연구이다 (Godard et al. 2019, Almalioglu et al. 2019). 마지막으로, 신경망 구조 자체를 개선하는 연구이다 (Wang et al. 2019).

2.3 준지도식 깊이 추정 기법

준지도식 깊이 추정 기법은 학습 데이터로 조밀한 깊이 정보를 활용하지는 않지만, 이미지 이외의 다른 정보를 활용하는 기법을 의미한다. 일반적으로는 라이다로부터 측정된 듬성듬성한 깊이 정보를 활용하여 학습하는 경우 혹은 스테레오 카메라 등을 활용하여 두 이미지 사이의 자세 정보를 활용하여 학습하는 경우로 나뉜다.

듬성듬성한 깊이 정보를 활용하는 경우, 지도식 기법에서 활용한 깊이 참값과 추정값의 거리 (Eq. (1))를 최소화하도록 신경망을 학습한다. 또한, 깊이가 참값이 제공된 픽셀에서만 추정되는 현상을 막기 위해, 비지도식 기법에서 활용하는 이미지 거리 (Eq. (6)) 역시 최소화한다.

스테레오 카메라를 활용하는 경우, 연속된 이미지가 아닌 스테레오 이미지를 이용한다는 점을 제외하고는 비지도식 기법과 동일한 방식으로 학습을 수행한다 (Eqs. (4-6)). 스테레오 카메라는 두 이미지 사이의 자세 정보가 고정되어 있으며 미리 측정할 수 있으므로, 자세 정보를 네트워크를 이용하여 추정하지 않고, 그 대신 기 측정된 두 이미지 사이의 자세 정보를 활용한다는 차이점이 있다.

Garg et al. (2016)은 스테레오 이미지를 활용한 준지도식 기법을 처음으로 제안하였다. 이를 통해 깊이 추정 신경망을 깊이 참값 없이도 학습이 가능하다는 것을 보였다.

Kuznietsov et al. (2017)은 스테레오 이미지 및 듬성듬성한 라이다 정보를 활용하여 준지도식 기법을 제안하였다. 이러한 준지도식 기법이 지도식 기법 및 스테레오 이미지만 활용하는 준지도식 기법에 비해 깊이 추정 정확도가 높다는 사실을 보였다.

Godard et al. (2017)은 스테레오 이미지를 활용한 준지도식 기법에서, 왼쪽 이미지를 통해 오른쪽 이미지의 깊이 정보 역시 추정하면서 이렇게 추정한 왼쪽/오른쪽 깊이 정보의 일관성을 부여하는 손실 함수를 추가하였다. 이를 통하여 깊이 추정의 정확도를 향상하였다.

Aleotti et al. (2018)은 스테레오 이미지를 활용한 준지도식 기법에 생성적 적대 신경망 기법을 적용하였다. 단순히 변환한 이미지와 측정된 이미지의 거리 (Eq. (6))를 최소화하는 대신, 변환한 이미지와 측정한 이미지를 구분하는 신경망을 설계하고, 깊이 추정 네트워크는 구분하는 신경망이 구분하지 못하도록 학습하는 방식이다. 이를 이용하여 깊이 추정 정확도를 향상하였다.

준지도식 학습 기법은 기본적으로는 추가 데이터를 어떻게 가공하는지에 초점을 맞추어 연구가 진행되어왔다. 크게는 스테레오 이미지만을 활용하는 경우 (Garg et al. 2016, Godard et al 2017, Aleotti et al. 2018), 듬성듬성한 라이다 정보를 활용하는 경우로 (Kuznietsov et al. 2017) 구분할 수 있다. 다만, 스테레오 이미지를 활용한 기법은 비지도식 기법과 동일한 방법론을 활용하므로, 최근에는 비지도식 기법을 개발하고 이에 스테레오 학습을 수행하는 경향이 있다 (Godard et al. 2019).

2.4 도메인 적응 기법

실제 환경에서는 학습을 위한 깊이 참값을 얻기가 어렵기 때문에 컴퓨터 그래픽스로 구현한 가상 환경에서 샘플링한 이미지와 이에 해당하는 깊이 참값을 이용하여 심층 신경망을 학습하는 방법이 고안되었다. 하지만, 가상 환경에서 모은 데이터를 활용하여 신경망을 학습할 경우 실제 환경에서 모델 성능이 급격히 떨어지는 것을 확인할 수 있는데 이는 사람의 눈으로 볼 때는 비슷한 환경일지라도, 실제 환경과 가상 환경에서 샘플링한 데이터의 픽셀 값 분포에서 매우 큰 차이가 존재하고 이 분포 차이, 혹은 도메인 차이로 인하여 성능 감소가 발생한다.

깊이 추정 문제를 해결하는데 있어서 도메인 적응 기법은 일반적으로 생성적 적대 신경망을 활용하는 경향이 있다. 가상 환경에서 샘플링 된 이미지를 실제 환경 도메인으로 매핑하는 신경망 G와 주어진 이미지가 실제 환경 이미지일 확률을 추정하는 신경망 D를 설계한다. 그런 다음, Eq. (8)과 같은 GAN loss를 활용하여 학습을 수행한다. 여기서 G는 Eq. (8)이 최소화되도록, D는 Eq. (8)이 최대화되도록 학습을 수행하게 된다.

$$L_{GAN} = \mathrm{E}_{x_s}\left[\log\left(\mathrm{D}\left(\mathrm{G}\left(x_s\right)\right)\right)\right] + \mathrm{E}_{x_t}\left[\log\left(1-\mathrm{D}\left(\mathrm{G}\left(x_t\right)\right)\right)\right] \quad (8)$$

여기서, $\mathrm{E}_x[\bullet]$는 x 분포에서 $\bullet$의 기대값, x_s는 샘플링된 이미지, x_t

는 실제 이미지이다.

Kundu et al. (2018)은 합성곱 신경망을 이용하여 실제 환경에서 샘플링한 이미지와 가상 환경에서 샘플링한 이미지에 대하여 각각 이미지 특징 벡터를 추출하고 두 벡터에 대하여 판별기를 도입해 실제 혹은 가상 환경 점수를 측정하여 두 특징 벡터의 차이를 줄이는 적대적 학습을 이용하여 도메인 차이로 인한 성능 감소를 최소화하였다.

Zheng et al. (2018)은 end-to-end 방식으로 도메인 적응과 깊이 추정을 동시에 학습하는 네트워크를 제안하였고, CycleGAN (Zhu et al. 2017)과 깊이 추정 네트워크가 서로의 성능을 상호 보완하며 함께 학습된다.

Zhao et al. (2019) 역시 도메인 적응과 깊이 추정을 동시에 학습하는 방식을 제안하였는데, 일방향 도메인 적응인 CycleGAN (Zhu et al. 2017)과 다르게 양방향으로 도메인을 전이할 수 있는 DiscoGAN을 채택하였고, 또한 스테레오 이미지와 손실 함수 (Eq. (6))를 이용하여 실제 환경에서의 깊이 추정 성능을 향상시켰다.

도메인 적응 학습 기법은 가상 환경에서 학습한 모델을 실제 환경에서도 비슷한 성능을 발휘하도록 연구가 진행되었다. 일반적으로 적대적 학습 신경망을 활용하여 가상 환경의 이미지를 실제 환경과 비슷하게 모사하였고 (Kundu et al. 2018, Zheng et al. 2018), 최근에는 실제 환경에서의 스테레오 이미지를 추가로 활용하여 가상 환경과 실제 환경에서 동시에 학습을 진행하는 연구 방법 또한 제안되었다 (Zhao et al. 2019).

3. 학습기반 깊이 추정 기법의 성능 검증

대부분의 깊이 추정 기법은 다른 기법과 비교 분석을 통하여 제안된 기법의 우수성을 입증한다. 이를 위해, 동일한 환경에서의 학습 및 검증이 필요하다. 이를 위하여, 대부분의 기법은 공개된 데이터셋을 활용하여 학습 및 검증을 수행한다. 또한, 다른 기법과의 비교 분석을 위하여 널리 사용되는 성능 지표를 제시한다.

3.1 데이터셋

학습기반 깊이 추정 기법 연구들은 동일한 환경에서의 성능을 비교, 검증을 수행하여 다른 알고리즘과의 차별성 및 우수점을 입증하기 위하여, 공개된 데이터셋을 활용하여 동일한 환경에서 성능을 제시한다. 이때, 공개된 데이터셋을 동일한 방식으로 학습용 데이터와 검증용 데이터를 나누어서, 동일한 성능 검증 환경을 구성하게 된다.

NYU Depth: 실내 깊이 추정 데이터셋으로는 Silberman et al. (2012)이 공개한 NYU Depth Dataset V2가 광범위하게 활용된다. 해당 데이터셋은 RGB-D 카메라를 활용하여 실내에서 약 1,500장의 이미지 및 조밀한 깊이 참값을 제공한다. 조밀한 깊이 참값이 제공되기에, 지도식 깊이 추정 기법을 연구하는데 있어서 보편적으로 활용되는 데이터셋이다. 학습용 데이터와 검증용 데이터는 해당 데이터셋에서 구분한 기준을 따른다.

KITTI: KITTI 데이터셋은 Geiger et al. (2012)이 공개한 데이터셋으로, 차량 주행 환경에서 수집한 데이터셋이다. 이 데이터셋은 스테레오 이미지, 6-dof 자세, 속도, 64채널 라이다 등의 정보를 제공한다. 라이다 측정값은 듬성듬성한 깊이 정보만을 제공하기에, 지도식 학습 기법을 연구하기 위해서는 라이다에서 측정된 듬성듬성한 깊이 정보를 기반으로 보간을 통해 조밀한 깊이 정보로 만들 필요가 있다. 이러한 이유로, 준지도식 혹은 비지도식 깊이 추정 기법을 연구하는데 보편적으로 활용되는 데이터셋이다.

깊이 추정에 있어서는 총 56개의 주행 데이터 중에서 28개는 학습용 데이터로 활용하고 28개는 검증용 데이터로 활용하는 Eigen split (Eigen et al. 2014)이 가장 보편적으로 활용된다. 비지도식 깊이 추정 기법은 상대 자세 정보 역시 추정하게 되는데, 추정한 상대 자세를 검증하는 경우 보편적으로 10개의 주행 데이터가 포함된 odometry 데이터셋을 활용한다. 이 때, 1-8번 주행 데이터는 학습용 데이터로 활용하고 9-10번 데이터는 검증용 데이터로 활용하는 것이 일반적이다.

Cityscapes: Cityscapes 데이터셋은 Cordts et al. (2016)이 공개한 데이터셋으로, 도시 내의 차량 주행 환경에서 수집한 데이터셋이다. 일반적으로는 50개의 객체에 대해 분할 (segmentation) 참값을 제공하기에, 인식 연구에 많이 활용된다. 깊이 추정 기법에서는 그 자체를 학습 및 검증 데이터로 활용하기 보다는, 앞선 KITTI 데이터셋에서 학습을 수행하는데 있어서 추가적인 학습용 데이터로 활용하는 경우가 많다.

3.2 검증 지표

정량적 성능 비교를 위하여 깊이 참값 d_i 및 추정된 깊이 $\hat{d}_i$에 대하여, 평균 제곱근 오차 (Root Mean Square Error; RMSE), RMSE log, 절대 상대 오차 (Absolute Relative Error; Abs Rel), 제곱근 상대 오차 (Square Relative Error; Sq Rel)가 많이 활용된다. 이 네 지표는 Eq. (9)와 같이 정의된다. 이 네 지표는 낮을수록 성능이 좋다는 의미이다.

$$\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|d_i-\hat{d}_i\right\|_2^2} \\
\text{RMSE log} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\log d_i-\log \hat{d}_i\right\|^2} \\
\text{Abs Rel} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left|d_i-\hat{d}_i\right|^2}{d_i} \\
\text{Sq Rel} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left\|d_i-\hat{d}_i\right\|_2^2}{d_i}
\end{aligned} \tag{9}$$

또한, 정확도 지표로 δ가 많이 활용된다. 정확도 δ는 Eq. (10)과 같이 정의되며, 일반적으로는 δ <1.25, δ <1.25^2, δ <1.25^{3}인 픽셀의 비율을 제시한다. 이 세 지표는 높을수록 성능이 좋다는 의미이다.

$$\delta = \max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) \tag{10}$$

Table 1. The depth estimation performance in NYU Depth dataset (Silberman et al. 2012). Train label means the training data of given methods: depth is depth ground-truth, and virtual is virtually generated dataset with image and dense depth ground-truth pairs. All results from each cited paper.

		Lower is better				Higher is better		
Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
Eigen et al. (2014)	depth	0.215	0.212	0.907	0.285	0.611	0.887	0.971
Laina et al. (2016)	depth	0.127	-	0.573	0.195	0.811	0.953	0.988
Fu et al. (2018)	depth	0.115	-	0.509	-	0.828	0.965	0.992
Wofk et al. (2019)	depth	-	-	0.604	-	0.771	-	-
Lee & Kim (2019)	depth	0.131	0.087	0.538	0.180	0.837	0.971	0.994
Kundu et al. (2018)	depth + virtual	0.114	-	0.506	-	0.856	0.966	0.991
Kundu et al. (2018)	virtual	0.136	-	0.603	0.057	0.805	0.948	0.982
Zheng et al. (2018)	virtual	0.157	0.125	0.556	0.199	0.779	0.943	0.983

Table 2. The depth estimation performance in KITTI dataset (Geiger et al. 2012) with eigen split (Eigen et al. 2014) and 80m cap. Train label means the training data of given methods: depth is depth ground-truth, mono/stereo seq is sequential data of mono/stereo, CS is additional training data from cityscape dataset (Cordts et al. 2016), stereo is stereo dataset with no sequential information, depth* is sparse depth ground-truth and virtual is virtually generated dataset with image and dense depth ground-truth pairs. All results from each cited paper.

		Lower is better				Higher is better		
Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
Eigen et al. (2014)	depth	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Fu et al. (2018)	depth	0.072	0.307	2.727	0.120	0.932	0.984	0.994
Wang et al. (2019)	depth	0.077	0.205	1.698	0.110	0.941	0.990	0.998
Zhou et al. (2017)	mono seq	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Zhou et al. (2017)	mono seq (CS)	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Yin & Shi (2018)	mono seq	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Yin & Shi (2018)	mono seq (CS)	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Godard et al. (2019)	mono seq	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Almalioglu et al. (2019)	mono seq	0.150	0.141	5.448	0.216	0.808	0.939	0.975
Almalioglu et al. (2019)	mono seq (CS)	0.138	1.155	4.412	0.232	0.820	0.939	0.976
Wang et al. (2019)	mono seq	0.112	0.418	2.320	0.153	0.882	0.974	0.992
Garg et al. (2016)	stereo	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Kuznietsov et al. (2017)	stereo + depth*	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Godard et al. (2017)	stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. (2017)	stereo (CS)	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Aleotti et al. (2018)	stereo	0.119	1.239	5.998	0.212	0.846	0.940	0.976
Aleotti et al. (2018)	stereo (CS)	0.098	0.908	5.164	0.177	0.879	0.961	0.986
Godard et al. (2019)	stereo seq	0.106	0.806	4.630	0.193	0.876	0.958	0.980
Godard et al. (2019)	stereo	0.107	0.849	4.764	0.201	0.874	0.953	0.977
Kundu et al. (2018)	depth + virtual	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Kundu et al. (2018)	virtual	0.214	1.932	7.157	0.295	0.665	0.882	0.950
Zheng et al. (2018)	depth + virtual	0.168	1.199	4.674	0.243	0.772	0.912	0.966
Zhao et al. (2019)	stereo + virtual	0.149	1.003	4.995	0.227	0.824	0.941	0.973

KITTI dataset과 같이 라이다를 이용한 깊이 참값을 활용하는 경우, 조밀한 깊이 참값의 경우 라이다 참값에서 보간한 참값을 활용하게 되고, 이는 참값 데이터의 왜곡을 가져올 수 있다. 이러한 상황을 방지하기 위하여 성능 검증에서는 조밀하지 않은, 라이다로부터 측정된 영역의 깊이 정보만을 비교하게 된다.

비지도식 학습 기법의 경우, 스케일 모호성으로 인하여 스케일이 추정되지 않는다는 문제점이 있다. 이를 검증하기 위하여 깊이 참값으로부터 스케일을 복원한 후에 Eqs. (9)와 (10)의 성능지표를 구한다. 일반적으로 깊이 참값과 추정된 깊이 각각의 중간값을 스케일이라고 가정하고, Eq. (11)과 같은 방식으로 스케일을 복원한다.

$$\tilde{d}_i = \lambda \hat{d}_i \quad \text{where } \lambda = \frac{\text{median}(d_i)}{\text{median}(\hat{d}_i)} \tag{11}$$

Tables 1과 2는 NYU Depth 데이터셋 및 KITTI 데이터셋에서의 여러 학습 기반 깊이 추정 알고리즘의 성능을 비교한 것이다. 기법이 발전해 나감에 따라, 깊이 추정 정확도가 점차 늘어나서 현재 KITTI 데이터셋 기준 약 1.7 m 평균 오차 수준에 도달한 것을 확인할 수 있다. 기본적으로 지도식 기법이 비지도식 및 준지도식 기법에 비해 뛰어난 성능을 보인다. 이는 비지도식/준지도식 기법은 실제 깊이 정보가 아닌 에피폴라 기하학에 기반하나, 이는 물체 가림, 동적 물체 등으로 인하여 오차 요인이 발생할 요인이 크기 때문이다.

한가지 유의해야 할 점은, 비지도식 기법은 Eq. (11)의 방식을 활용하여 프레임 단위로 깊이 참값의 스케일을 복원하게 되므로, 지도식 및 준지도식 기법과의 직접적인 성능지표 비교는 왜곡된 결과를 가져올 수 있다는 점이다. 예를 들어, Godard et al. (2019)에서 단안 영상 이미지 (mono seq)를 활용한 비지도식 기법의 정확도(δ)가 스테레오 영상 이미지 (stereo seq)를 활용한 준지도식 기법보다, 더 적은 학습 데이터를 활용하였음에도 높은 정확도를 달성한 것을 확인할 수 있다. 이는 단순히 비지도식 기법에 한하여 Eq. (11)를 활용하여 스케일을 복원하였기 때문이다.

Fig. 2는 지도식, 비지도식, 준지도식 대표적인 기법의 깊이

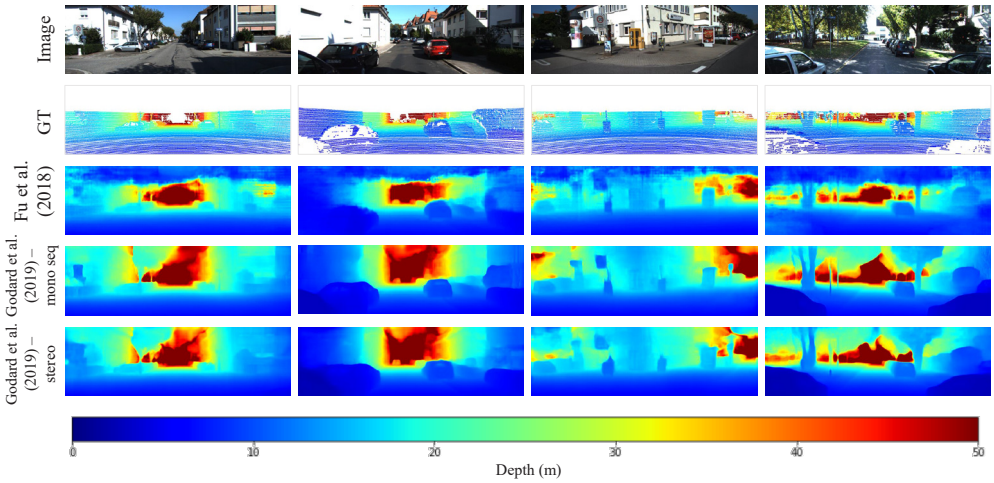

Fig. 2. The example of the monocular depth estimation. Each result is calculated from author provided code and weights. For unsupervised method (Godard et al. 2019) with mono seq, we recover scale from ground-truth depth as in Eq. (11).

추정 결과의 예시를 나타낸 것이다. 비지도식 기법은 깊이 추정을 수행함에 있어서 에피폴라 기하학에 의존함과 동시에, 에피폴라 기하학을 구성하기 위해서 카메라가 움직이게 된다. 이러한 이유로, 에피폴라 기하학을 구성하는 과정에서 필연적인 시차가 발생하게 되고 이는 일부 물체의 깊이 추정 왜곡 현상을 불러올 수 있다.

4. 결론

본 논문에서는 학습 기반 깊이 추정 알고리즘을 학습 데이터에 따라 지도식, 비지도식, 준지도식 방식으로 분류하고, 해당 기법 및 깊이 추정 알고리즘에 대한 기본적인 설명 및 현존하는 연구들이 깊이 추정 성능을 향상시키기 위하여 어떠한 방식을 도입하였는지를 논의하였다.

ACKNOWLEDGMENTS

본 연구는 과학기술정보통신부의 재원으로 한국연구재단, 무인이동체원천기술개발사업단의 지원을 받아 무인이동체원천기술개발사업을 통해 수행되었음 (NRF-2020M3C1C1A01086411).

AUTHOR CONTRIBUTIONS

Conceptualization, C. L., D. S., and H. K.; investigation, C. L., and D. S.; writing—original draft preparation, C. L, D. S..; writing—review and editing, D. S. and H. K.; visualization, C. L.; supervision, H. K.; project administration, H. K.; funding acquisition, H. K.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aleotti, F., Tosi, F., Poggi, M., & Mattoccia, S. 2018, Generative adversarial networks for unsupervised monocular depth prediction, in Proceedings of the European Conference on Computer Vision Workshops, Munich, Germany, Sep 2018, pp.337-354. https://doi.org/10.1007/978-3-030-11009-3_20

Almalioglu, Y., Saputra, M. R. U., de Gusmao, P. P. B., Markham, A., & Trigoni, N. 2019, Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks, in International Conference on Robotics and Automation, Montreal, QC, May 2019, pp. 5474-5480. https://doi.org/10.1109/ICRA.2019.8793512

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., et al. 2016, The cityscapes dataset for semantic urban scene understanding, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, Jun 2016, pp.3213-3223. https://doi.org/10.1109/CVPR.2016.350

Eigen, D., Puhrsch, C., & Fergus, R. 2014, Depth map prediction from a single image using a multi-scale deep network, in Advances in Neural Information Processing Systems, Cambridge, MA, Dec 2014, pp.2366-2374. https://dl.acm.org/doi/10.5555/2969033.2969091

Forster, C., Pizzoli, M., & Scaramuzza, D. 2014, SVO: Fast semi-direct monocular visual odometry, in IEEE international conference on robotics and automation,

Hong Kong, China, Jun 2014, pp.15-22. https://doi.org/10.1109/ICRA.2014.6906584

Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. 2018, Deep Ordinal Regression Network for Monocular Depth Estimation, in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, Jun 2018, pp.2002-2011. https://doi.org/10.1109/CVPR.2018.00214

Garg, R., Bg, V. K., Carneiro, G., & Reid, I. 2016, Unsupervised Cnn for Single View Depth Estimation: Geometry to the Rescue, in European conference on computer vision, Amsterdam, the Netherlands, Oct 2016, pp.740-756. https://doi.org/10.1007/978-3-319-46484-8_45

Geiger, A., Lenz, P., & Urtasun, R. 2012, Are we ready for autonomous driving? the kitti vision benchmark suite, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Providence, RI, Jun 2012, pp.3354-3361. https://doi.org/10.1109/CVPR.2012.6248074

Godard, C., Aodha, O. M., & Brostow, G. J. 2017, Unsupervised Monocular Depth Estimation with Left-Right Consistency, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 2017, pp.270-279. https://doi.org/10.1109/CVPR.2017.699

Godard, C., Aodha, O. M., Firman, M., & Brostow, G. J. 2019, Digging Into Self-Supervised Monocular Depth Estimation, in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, Oct 2019, pp.3828-3838. https://doi.org/10.1109/ICCV.2019.00393

He, K., Zhang, X., Ren, S., & Sun, J. 2016, Deep Residual Learning for Image Recognition, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 27-30 Jun 2016, pp.770-778. https://doi.org/10.1109/CVPR.2016.90

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., et al. 2017, Mobilenets: Efficient convolutional neural networks for mobile vision applications, Jun 9, Retrieved from https://arxiv.org/abs/1704.04861

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. 2015, in Advances in Neural Information Processing Systems, Montreal, CA, Dec 2015, pp.2017-2025. https://dl.acm.org/doi/abs/10.5555/2969442.2969465

Kundu, J. N., Uppala, P. K., Pahuja, A., & Babu, R. V. 2018, AdaDepth: Unsupervised Content Congruent Adaptation for Depth Estimation, in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18-23 Jun 2018, pp.2656-2665. https://doi.org/10.1109/CVPR.2018.00281

Kuznietsov, Y., Stuckler, J., & Leibe, B. 2017, Semi-Supervised Deep Learning for Monocular Depth Map Prediction, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 21-26 July 2017, pp.6647-6655. https://doi.org/10.1109/CVPR.2017.238

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. 2016, Deeper Depth Prediction with Fully Convolutional Residual Networks, in Fourth International Conference on 3D Vision, Stanford, CA, 25-28 Oct 2016, pp.239-248. https://doi.org/10.1109/3DV.2016.32

Lee, J. & Kim, C. 2019, Monocular Depth Estimation Using Relative Depth Maps, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 15-20 Jun 2019, pp.9729-9738. https://doi.org/10.1109/CVPR.2019.00996

Mur-Artal, R., Montiel, J. M. M., & Tardos, J. D. 2015, ORB-SLAM: a Versatile and Accurate Monocular SLAM system, IEEE transactions on robotics, 31, 1147-1163. https://doi.org/10.1109/TRO.2015.2463671

Ranftl, R., Vineet, V., Chen, Q., & Koltun, V. 2016, Dense Monocular Depth Estimation in Complex Dynamic Scenes, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 27-30 Jun 2016, pp.4058-4066. https://doi.org/10.1109/CVPR.2016.440

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R., 2012, Indoor Segmentation and Support Inference from RGBD Images, in European Conference on Computer Vision, Firenze, Italy, Oct 2012, pp.746-760. https://doi.org/10.1007/978-3-642-33715-4_54

Wang, R., Pizer, S. M., & Frahm, J. 2019, Recurrent Neural Network for (Un-)Supervised Learning of Monocular Video Visual Odometry and Depth, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, Jun 2019, pp.5555-5564. https://doi.org/10.1109/CVPR.2019.00570

Wofk, D., Ma, F., Yang, T. J., Karaman, S., & Sze, V. 2019, FastDepth: Fast Monocular Depth Estimation on Embedded Systems, in International Conference on Robotics and Automation, Montreal, QC, 20-24 May 2019, pp.6101-6108. https://doi.org/10.1109/ICRA.2019.8794182

Yin, Z. & Shi, J. 2018, GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose, in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18-

23 Jun 2018, pp.1983-1992. https://doi.org/10.1109/CVPR.2018.00212

Zhao, S., Fu, H., Gong, M., & Tao, D. 2019, Geometry-Aware Symmetric Domain Adaptation for Monocular Depth Estimation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 15-20 Jun 2019, pp.9788-9798. https://doi.org/10.1109/CVPR.2019.01002

Zheng, C., Cham, T. J., & Cai, J. 2018, T2net: Synthetic-to-Realistic Translation for Solving Single-Image Depth Estimation Tasks, in Proceedings of the European Conference on Computer Vision, Munich, Germany, 8-14 Sep 2018, pp.798-814. https://doi.org/10.1007/978-3-030-01234-2_47

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. 2017, Unsupervised Learning of Depth and Ego-Motion from Video, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 21-26 July 2017, pp.1851-1858. https://doi.org/10.1109/CVPR.2017.700

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. 2017, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22-29 Oct 2017, pp.2223-2232. https://doi.org/10.1109/ICCV.2017.244

Chungkeun Lee received the B.S. degree in Mechanical and Aerospace Engineering in 2014. He is currently pursuing the Ph.D. degree in Mechanical and Aerospace Engineering at Seoul National University. His research interests are robotic applications with deep-visual learning including perception and depth estimation.

Dongseok Shim received the B.S. degree in Mechanical and Aerospace Engineering from Seoul National University in 2020. He is currently pursuing M.S. degree in Aerospace Engineering from Seoul National University. His research interests include machine learning, visual perception, and the learning-based application in robotics.

H. Jin Kim received the B.S. degree from Korea Advanced Institute of Technology (KAIST) in 1995, and the M.S. and Ph.D. degrees in Mechanical Engineering from University of California, Berkeley, in 1999 and 2001, respectively. From 2002 to 2004, she was a Postdoctoral Researcher in Electrical Engineering and Computer Science, UC Berkeley. In 2004, she joined the Department of Mechanical and Aerospace Engineering at Seoul National University as an Assistant Professor, where she is currently a Professor. Her research interests include intelligent control of robotic systems and motion planning.