KIPS Trans. Softw. and Data Eng.
Vol.11, No.12 pp.509~516
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.12.509

Predicting Unseen Object Pose with an Adaptive Depth Estimator

Sungho Song† · Incheol Kim††

ABSTRACT

Accurate pose prediction of objects in 3D space is an important visual recognition technique widely used in many applications such as scene understanding in both indoor and outdoor environments, robotic object manipulation, autonomous driving, and augmented reality. Most previous works for object pose estimation have the limitation that they require an exact 3D CAD model for each object. Unlike such previous works, this paper proposes a novel neural network model that can predict the poses of unknown objects based on only their RGB color images without the corresponding 3D CAD models. The proposed model can obtain depth maps required for unknown object pose prediction by using an adaptive depth estimator, AdaBins,. In this paper, we evaluate the usefulness and the performance of the proposed model through experiments using benchmark datasets.

Keywords : 3D Vision, Unknown Object, 6D Pose Prediction, Depth Estimation, Deep Neural Network

적응형 깊이 추정기를 이용한 미지 물체의 자세 예측

송 성 호† · 김 인 철††

요 약

3차원 공간에서 물체들의 정확한 자세 예측은 실내외 환경에서 장면 이해, 로봇의 물체 조작, 자율 주행, 증강 현실 등과 같은 많은 응용 분야들에서 폭넓게 활용되는 중요한 시각 인식 기술이다. 물체들의 자세 예측을 위한 과거 연구들은 대부분 각 인식 대상 물체마다 정확한 3차원 CAD 모델을 요구한다는 한계점이 있었다. 이러한 과거 연구들과는 달리, 본 논문에서는 3차원 CAD 모델이 없어도 RGB 컬러 영상들만 이용해서 미지 물체들의 자세를 예측해낼 수 있는 새로운 신경망 모델을 제안한다. 제안 모델은 적응형 깊이 추정기인 AdaBins를 이용하여 스스로 미지 물체 자세 예측에 필요한 각 물체의 깊이 지도를 효과적으로 추정해낼 수 있다. 벤치마크 데이터 집합들을 이용한 다양한 실험들을 통해, 본 논문에서 제안한 모델의 유용성과 성능을 평가한다.

키워드 : 3D 비전, 미지 물체, 6D 자세 예측, 깊이 추정, 심층 신경망

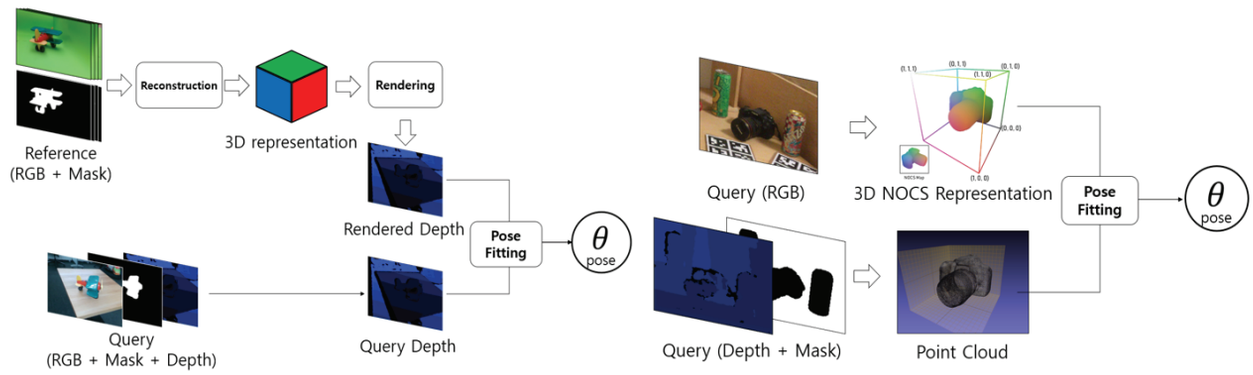
1. 서 론

3차원 공간에서의 물체 탐지(object detection)와 자세 예측(pose prediction)은 실내외 환경에서 장면 이해(scene understanding), 로봇의 물체 조작(robotic manipulation), 자율 주행(autonomous driving), 증강 현실(augmented reality) 등과 같은 다양한 응용 분야들에서 폭넓게 활용되는 중요한 시각 인식 기술이다. 특히 이 중에서 3차원 공간에서 물체의 6D 자세 예측은 카메라를 중심으로 특정 물체의 3축 회전(rotation)과 3축 변환(translation)를 알아내는 기술이다. 따라서 일반적으로 물체의 6D 자세 예측은 해당 물체를 둘러싸는 직육면체 형태의 경계 상자(bounding box)를 알아내려는 3차원 물체 탐지보다 더 높은 정밀도를 요구하는 작업이다.

물체의 6D 자세 예측에 관한 과거 연구들은 대부분 대상 물체의 정확한 3차원 CAD 모델을 이용하는 개체-수준 자세 예측(instance-level pose prediction) 방식을 채택하였다. 최근에 와서는 이러한 개체-수준의 자세 예측기들은 매우 높은 수준의 자세 정확도를 얻는 데 성공하였으나, 인식 대상 물체마다 모두 3차원 CAD 모델이 확보되어야만 자세 예측이 가능하다는 한계는 뛰어넘지 못하고 있다 [1]. 반면에, 최근 들어서는 이러한 개체-수준의 자세 예측기들의 한계성을 극복하기 위해, 인식 대상 물체가 속한 범주(category)나 동일 범주의 다른 개체들의 3차원 표현은 알 수 있으나 해당 물체의 3차원 CAD 모델은 가지고 있지 않다고 가정하는 미지 물체(unseen object)에 관한 범주-수준의 자세 예측(category-level pose prediction)에 관한 연구가 활발하다[2-4].

기존 연구들에서 제시된 미지 물체에 대한 대표적인 범주-수준의 자세 예측 방식들로는 (a) 3차원 재건과 랜더링(3D re-

※ 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구과제(No. 2020-0-00096. 클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발)입니다. 또한, 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0008691, 2022년 산업혁신인재성장지원사업)
※ 이 논문은 2022년 한국정보처리학회 ASK 2022의 우수논문으로 "단안 카메라 깊이 추정기를 이용한 미지 물체의 자세 추정"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터공학부 교수
Manuscript Received : July 21, 2022
First Revision : September 13, 2022
Accepted : September 18, 2022
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)



(a) 3D reconstruction and rendering (b) 3D NOCS representation for each category

Fig. 1. Two Different Kinds of Unseen Object Pose Prediction

construction and rendering) 방식과 (b) 범주별 3차원 NOCS 표현(3D NOCS representation for each category)을 이용하는 방식 등이 있다. 3차원 재건과 랜더링 방식[2, 13, 14]은 Fig. 1 (a)와 같이 다수의 다른 물체 데이터들로 학습된 신경망을 이용해, 인식 대상 물체에 관한 소량의 참조 영상(reference image)들로부터 해당 물체의 3차원 표현을 재건한 뒤, 이를 랜더링함으로써 2차원의 깊이 지도(depth map)를 추정한다. 그리고 대상 물체의 입력 깊이 지도와 추정된 깊이 지도 간의 매칭 과정을 통해, 해당 물체의 6D 자세를 예측해낸다.

반면에, 범주별 3차원 NOCS(Normalized Object Coordinate Space) 표현을 이용하는 방식[3, 4]은 Fig. 1 (b)와 같이 인식 대상 물체에 관한 별도의 참조 영상 없이 입력 RGB 영상과 깊이 지도로부터 각각 해당 물체가 속한 범주의 표준 NOCS 표현과 해당 물체의 포인트 클라우드(point cloud)를 생성한 다. 그리고 이 2개의 3차원 표현 간의 매칭 과정을 통해, 해당 물체의 자세를 예측한다.

하지만, 앞서 언급한 기존의 두 가지 미지 물체 자세 예측 방식 모두 인식 대상 물체의 깊이 지도를 입력 데이터로 요구한다. 따라서 물체 자세 예측을 위해서는 일반 RGB 카메라 외에 추가로 적외선(IR)이나 라이다(Lidar) 등 깊이 측정 센서가 필요하다. 한편 최근에는 RGB 단안 카메라 영상으로부터 깊이 지도를 추정해내는 단안 카메라 깊이 추정(monocular depth estimation) 기술이 급속히 발전하여 비교적 높은 성능을 보여주고 있다[5, 6].

본 논문에서는 RGB 컬러 영상만을 이용해 미지 물체들의 자세를 추정해낼 수 있는 새로운 범주-수준 자세 예측 신경망 모델을 제안한다. 기존의 미지 물체 자세 예측 모델들과는 달리, 특히 제안 모델에서는 적응형 단안 카메라 깊이 추정기인 AdaBins를 이용함으로써 스스로 물체 자세 예측에

필요한 물체의 깊이 지도를 구해낼 수 있다. 따라서 제안 모델은 미지 물체의 자세 예측을 위해 3D CAD 모델도, 깊이 지도도 요구하지 않는 매우 높은 고수준의 사용자 편의성을 제공한다. 본 논문에서는 NYU-Depth-v2[6]와 REAL-275[3] 등 대규모 벤치마크 데이터 집합들을 이용한 정량 및 정성 평가 실험들을 통해, 제안 모델의 유용성과 성능을 평가한다. 본 논문의 2장에서는 관련 선행 연구들을 살펴보고, 3장에서는 제안 모델의 설계에 관해 설명한다. 4장에서는 제안 모델의 구현과 성능 분석 실험 결과들에 대해 설명하고, 5장에서는 결론과 향후 연구를 정리한다.

2. 관련 연구

2.1 3차원 물체의 6D 자세 예측

3차원 물체의 6D 자세 예측에 관한 기존 연구들은 대부분 개체 수준의 자세 예측 방식[7-12]에 주로 집중되었지만, 최근 들어서는 3차원 CAD 모델을 가지고 있지 않은 미지 물체들에 대한 범주 수준의 자세 예측 방식에 관한 연구도 활발히 소개되고 있다[2-4, 13, 14]. 개체 수준의 자세 예측에 관한 기존 연구들은 크게 RGB 영상만을 입력 데이터로 이용하는 연구들[7-10]과 깊이 지도를 포함한 RGB-D 영상을 이용하는 연구들[11, 12]로 다시 나눌 수 있다.

RGB 영상만을 사용하는 연구들 중에는 RGB 영상에서 키포인트(keypoint)들을 찾아내고 이것들을 활용해서 물체의 자세 예측에 이용하는 방식[7, 8]과 회귀(regression)를 통해 직접 물체의 자세를 추정하는 방식[9, 10]들이 있었다. [7, 8]의 연구들에서는 RGB 영상으로부터 찾아낸 키포인트들을 물체의 3차원 CAD 모델과 매치시켜 물체의 자세를 예측하였다. 한편, [9, 10]의 연구들에서는 합성곱 신경망(Convolutional Neural Network, CNN)을 통해 RGB 영상에서 시각적 특징 지도를 추출한 뒤, 다중 퍼셉트론 신경망(Multi-Layer Perceptron, MLP)을 통해 물체의 3D 위치와 3D 변환을 직접 예측하였다. 하지만 이러한 RGB 영상 기반의 물체 자세 예측 모델들은 물체 간에 폐색이 존재하거나 어수선한 장면에서는 자세 예측의 정확도가 낮아지는 문제점이 있었다.

한편, RGB-D 영상을 이용하는 연구들[11, 12]에서는 물체의 깊이 정보를 추가적으로 활용함으로써, 자세 예측 정확도를 개선하고자 하였다. [11]의 모델은 입력으로 주어지는 RGB 영상뿐만 아니라 깊이 지도에도 각각 합성곱 신경망(CNN)을 적용하여 상호보완적인 특징 지도들을 추출하였다. 그리고 이들

을 결합하여 하나의 멀티 모달 특징 지도를 얻고 이를 토대로 물체의 3D 위치와 3D 변환을 직접 예측하였다. 반면에 [12]의 모델은 RGB 영상에는 합성 곱 신경망을 적용하여 시각적 특징 지도(visual feature map)를 추출하였으나, 깊이 지도는 대응되는 포인트 클라우드를 생성한 후 3차원 포인트 특징 추출기인 PointNet 신경망을 적용하여 기하학적 특징 지도(geometric feature map)를 추출하였다. 이렇게 추출된 RGB 영상 기반의 시각적 특징 지도와 깊이 지도 기반의 기하학적 특징 지도를 서로 결합하여 물체의 자세 예측에 이용되는 멀티 모달 특징 지도를 생성하였다. 하지만 이러한 다양한 개체 수준의 물체 자세 예측 방식들은 모두 물체의 각 개체마다 정확한 CAD 모델과 추가 학습에 대한 요구 부담 때문에 폭넓게 활용되기 어려운 측면이 있다. 이러한 한계성을 벗어나고자, 새로운 범주 수준의 물체 자세 예측 모델들이 제안되기 시작되었다.

범주 수준의 물체 자세 예측을 위한 기존 연구들에서는 각 개체(instance)별 3차원 모델 대신 각 개체가 속한 범주(category)별로 해당 범주의 모든 개체들이 공유할 수 있는 공통의 3차원 표현을 활용하였다. 기존의 범주 수준 물체 자세 예측 연구들은 크게 3차원 재건과 랜더링 방식[2, 13, 14]과 범주별 3차원 NOCS 표현을 이용하는 방식[3, 4]으로 나눌 수 있다. [2, 13, 14]의 연구들에서는 자세를 인식하고자 하는 물체에 대응되는 잠재 공간(latent space)상의 3차원 표현을 재건한 후, 이것을 랜더링하여 해당 물체의 2차원 영상을 구한다. 그리고 물체의 랜더링 영상과 입력 영상을 비교함으로써, 해당 물체의 자세를 예측하려고 시도하였다. [2]의 모델은 물체에 관한 몇 장의 RGB 참조 영상들을 이용해 잠재 공간상의 3차원 표현을 재건하였고, 이것을 랜더링하여 물체 자세 예측에 이용할 예측된 깊이 지도(estimated depth map)들을 구하였다. 반면에 [13]의 모델은 물체의 깊이 지도와 물체 범주 정보를 이용해 잠재 공간상의 3차원 표현을 재건하였고, [2]의 모델과 같이 이것을 랜더링하여 물체 자세 예측에 이용할 예측된 깊이 지도를 생성하였다. 한편, VAE(Variational Auto-Encoder) 프레임워크에 기초한 [14]의 모델은 잠재 공간상의 3차원 표현을 토대로 원하는 자세의 물체 영상을 생성해줄 수 있는 자세-인식 영상 생성기(pose-aware image generator) 모듈을 학습시킨 뒤, 이 영상 생성기가 생성한 물체 영상과 실제 입력 영상과 비교함으로써 해당 물체의 자세를 예측하였다. 이와 같은 3차원 재건과 랜더링에 기초한 물체 자세 예측 모델들은 잠재 공간상의 3차원 표현을 얻기 위해, 공통적으로 해당 물체의 RGB 영상 외에 별도의 깊이 지도 혹은 깊이 지도 예측 모듈을 요구한다.

범주 수준의 물체 자세 예측을 위한 또 다른 연구들[3, 4]은 물체가 속한 범주에 대한 3차원 NOCS(Normalized Object Coordinate Space) 표현을 물체 자세 예측에 이용하였다. [3]의 모델은 물체의 RGB 영상과 깊이 지도로부터 해당 물체가 속한 범주의 표준 NOCS 표현과 동시에 해당 물체의 포인트 클라우드(point cloud)를 생성한 뒤, 이 2개의 3차원 표현을 서로 비교함으로써 해당 물체의 자세를 예측하였다. 한편, [4]의 모델은 범주의 NOCS 표현을 해당 물체의 NOCS 표현으로 변환해주는 변환 신경망(deformation network) 모듈을 별도로 활용하였다. 이 모듈은 물체의 RGB 영상과 깊이 지도 외에 해당 물체가 속한 범주의 3차원 NOCS 표현을 입력받아 해당 물체의 특성을 반영한 NOCS 표현을 생성하였다. 이와 같이 3차원 NOCS 표현을 이용하는 범주 수준의 물체 자세 예측 연구들도 모두 해당 물체의 RGB 영상 외에 별도의 깊이 지도를 입력으로 요구한다.

2.2 단안 카메라 깊이 추정

단안 카메라를 활용한 깊이 추정 기법은 단일 RGB 영상으로만 깊이 추정이 이루어진다. 단안 카메라 깊이 추정에 관한 기존 연구들은 학습 방법에 따라서 크게 지도 학습(supervised learning) 기법[15-19, 6]과 비지도 학습(unsupervised learning) 기법[20, 21]으로 나누어 볼 수 있다. 지도 학습 기반의 깊이 추정 방식들은 RGB 영상과 이에 대응되는 정답 깊이 지도(ground truth depth map)의 쌍들을 깊이 추정 모델을 위한 훈련 데이터로 활용하였다. 이때 정답 깊이 지도는 실내 환경에서는 주로 적외선(IR) 센서를, 실외 환경에서는 라이다(LIDAR) 센서를 이용해 확보하였다.

지도 학습 기반의 초기 깊이 추정 연구들은 대부분 입력 RGB 영상으로부터 이것에 대응되는 깊이 지도를 직접 생성하기 위해 합성곱 신경망(CNN)을 이용하였다[15, 16]. [15]의 연구에서는 개괄 합성 곱 신경망(coarse CNN)과 미세 합성 곱 신경망(refine CNN)으로 구성된 2단계의 계층적 신경망 구조를 제안하였다. 이 모델에서 개괄 합성 곱 신경망(coarse CNN)은 입력 RGB 영상에 대해 전역적 관점으로 개괄적인 깊이 지도를 예측하는 데 반해, 미세 합성 곱 신경망(refine CNN)은 개괄적인 깊이 지도를 토대로 지역적 세부 특성까지 고려해 정밀한 깊이 지도를 생성하는 역할을 수행하였다. 합성 곱 계층과 풀링(pooling) 계층들로만 구성된 [15]의 모델은 입력 RGB 영상에 비해 약 1/4의 낮은 해상도를 갖는 깊이 지도를 생성할 수밖에 없었다. 이러한 한계를 극복하고자, [16]의 모델은 역 합성 곱(Up Convolution) 계층들을 추가하여 입력 RGB 영상에 비해 약 1/2 정도의 비교적 고해상도 깊이 지도를 생성하였다.

이러한 합성 곱 신경망을 이용하는 초기 모델들에 이어 인코더-디코더(encoder-decoder) 신경망 구조를 이용해 고해상도 깊이 지도를 얻고자 하는 새로운 모델들이 활발히 제안되었다[17-19, 6]. [17]의 DenseDepth 모델은 표준 인코더-디코더 신경망 구조를 따르지만, ImageNet 데이터 집합으로 사전 학습시킨 DenseNet-169 신경망 모듈을 인코더로 채용하고 새로운 데이터 증강(data augmentation) 기법과 손실 함수(loss function)를 적용함으로써, 고해상도, 고품질의 깊이 지도를 생성하려고 하였다. 반면에 [18]의 DAV 모델은 표준 인코더와 디코더 사이에 비-지역적 깊이-주의집중 모듈을 추가하여 깊이-주의집중 볼륨(Depth-Attention Volume, DAV)을 생성하였다. 깊이-주의집중 볼륨은 동일 표면 점들(coplanar point) 간의 깊이 의존성을 효과적으로 표현하게 되며, DAV 모델은 이 정보를 추가로 이용함으로써 깊이 추정의 정확도를 높이고자 하였다. 한편, [19]의 BTS 모델은 고해상도, 고품질의 깊이 지도를 추정해내기위해, 디코더 블록마다 로컬 평면 유도

계층(local planar guidance layer)이 삽입된 새로운 디코더 구조를 이용하였다.

한편 [6]의 AdaBins 모델은 표준 인코더-디코더 블록의 후단에 깊이 구간(depth bin)들을 적응적으로 결정하기 위한 AdaBins 블록을 추가하였다. 이 모델에서 인코더-디코더 블록은 입력 RGB 영상에서 최종 깊이 지도가 아닌 다채널의 디코딩된 특징 지도(decoded feature map)를 출력으로 생성하고, Transformer 구조에 기초한 AdaBins 블록은 이 디코딩된 특징 지도에 맞추어 최종 깊이 지도 생성에 필요한 각 깊이 구간 중심과 폭들을 결정하는 역할을 수행한다. 이 모델은 깊이 구간 중심들의 선형 조합(linear combination)으로 최종 깊이 지도를 생성한다. 따라서 이 AdaBins 모델은 각 입력 RGB 영상별 특성에 맞추어 보다 정확도가 높은 깊이 지도를 생성할 수 있다는 장점이 있다.

지도 학습 기반의 깊이 추정을 위해서는 정답 깊이 지도(ground truth depth map)가 필요한데, 정답 깊이 지도 수집은 고가의 센서와 시간, 노력 등 고비용이 요구된다. 따라서 이것에 대한 하나의 대안으로서, 정답 깊이 지도를 요구하지 않는 비지도 학습 기반의 깊이 추정 방법에 관한 연구들도 등장하였다[20-23]. 비지도 학습 기반의 깊이 추정 모델들은 서로 다른 시점들(viewpoints)에서 촬영된 RGB 영상들이 주어지면, 에피폴라 기하학(epipolar geometry)을 적용하여 RGB 영상에 대응되는 깊이 지도를 추정할 수 있다는 점을 활용하였다[22,23]. 특히 [20]의 SfMLearner 모델은 카메라 포즈를 별도의 입력으로 요구하지 않는 대신, 연속된 RGB 비디오로부터 단일 시점 기반의 깊이 지도 추정과 다중 시점 기반의 카메라 포즈 추정을 동시에 수행하였다. 또한, [21]의 GeoNet 모델은 연속적인 RGB 비디오로부터 깊이 지도와 카메라 포즈뿐만 아니라, 광 흐름(optical flow)으로 표현되는 모션까지 함께 추정하도록 설계되었다. 이와 같은 비지도 학습 기반의 깊이 추정 모델들은 정답 깊이 지도를 요구하지 않는 편의성은 있으나, 아직 이들의 성능은 고해상도, 고품질의 밀집 깊이 지도를 보장할 수 없어서, 서비스 로봇과 같은 실내 환경 응용 분야보다는 자율 주행과 같은 실외 환경 응용 분야들에서 주로 활용을 찾고 있는 실정이다.

3. 미지 물체 자세 예측 모델

3.1 모델 개요

단일 RGB 입력 영상을 기반으로 미지 물체의 6D 자세 예측을 효과적으로 수행하기 위해 본 논문에서 제안하는 모델의 구성은 Fig. 2와 같다. 제안 모델 역시 미지 물체의 6D 자세 예측을 위해 범주별 3차원 NOCS 표현을 이용한다. 하지만 기존 모델들과는 달리, 별도의 깊이 추정 모듈을 채용해 깊이 지도를 자체적으로 생성한다. 제안 모델은 크게 (1) 깊이 추정(Depth Estimation) 모듈, (2) NOCS 지도 예측(NOCS Map Prediction) 헤드가 추가되어 새롭게 확장된 Mask-RCNN 신경망 모듈, (3) 3차원 확장(3D Lift-Up) 모듈, (4) 자세 추론(Pose Fitting) 모듈들로 구성된다. 깊이 추정 모듈은 RGB 입력 영상으로부터 그것에 대응하는 깊이 지도를 예측해낸다. 확장된 Mask-RCNN 신경망 모듈은 합성 곱 신경망(Convolutional Neural Network, CNN)을 통해 RGB 입력 영상으로부터 시각적 특징 지도(visual feature map)를 추출한 후, 관심 영역 제안 망(Region Proposal Network, RPN)을 통해 영상 내의 관심 영역들을 구한다. 그리고 각 관심 영역별로 물체의 종류(class), 물체의 경계 상자(bounding box, bbox), 물체의 마스크(mask), 물체의 NOCS 지도(NOCS map) 등을 예측한다. 3차원 확장 모듈은 물체 마스크를 기초로, (1) NOCS 지도와 결합하여 해당 물체의 3차원 NOCS 표현을 구하기도 하고, (2) 깊이 지도와 결합하여 해당 물체의 3차원 포인트 클라우드(point cloud)를 얻기도 한다. 마지막으로 자세 추론 모듈에서는 각 물체의 3차원 NOCS 표현과 포인트 클라우드를 서로 매칭함으로써, 해당 물체의 6D 자세와 크기를 예측한다.

3.2 깊이 추정

본 논문에서 제안하는 미지 물체 자세 예측 모델에서는 대표적인 감독 학습 기반의 단안 카메라 깊이 추정기인 AdaBins[6]를 이용한다. AdaBins 깊이 추정기는 Fig. 3과 같이 표준 인코더-디코더 합성곱 신경망(encoder-decoder convolutional

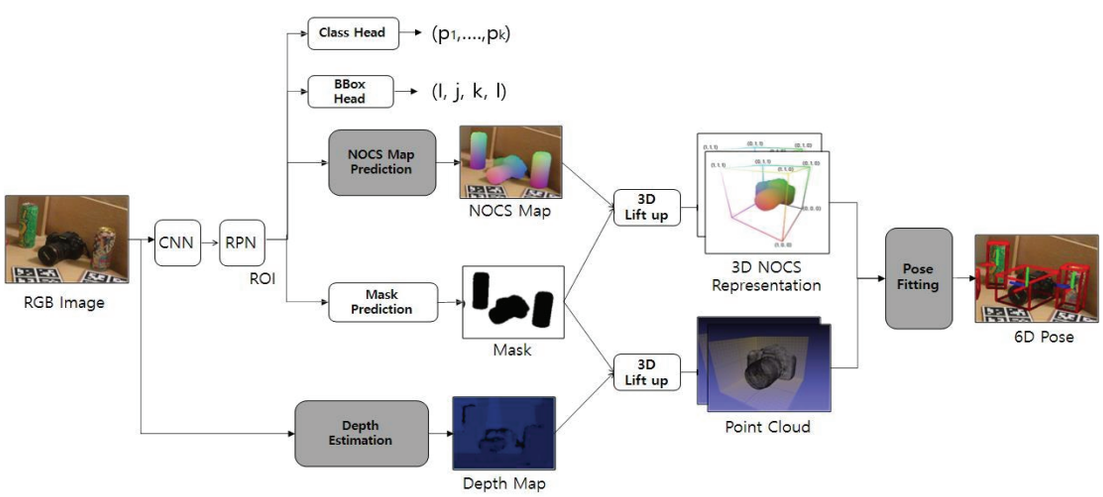


Fig. 2. Organization of the Proposed Model

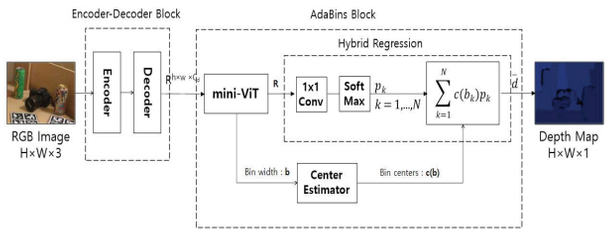


Fig. 3. Component Modules of the Depth Estimator

neural network) 구조에, 전역적 정보 처리를 위해 트랜스포머(Transformer) 기반 신경망 블록을 추가하였다. 트랜스포머 기반 신경망 블록인 AdaBins 블록은 깊이 범위(depth range)를 다수의 구간들(bins)로 나누며, 각 구간의 중심 값은 입력 영상에 맞게 적응적으로 추정된다. 따라서 AdaBins 깊이 추정기는 불충분한 전역 정보 처리로 인해 깊이 추정의 품질이 낮았던 기존의 깊이 추정기들의 단점을 극복하고 높은 깊이 추정 성능을 보여주고 있다.

Fig. 3과 같이, AdaBins 깊이 추정기의 인코더-디코더 블록은 RGB 입력 영상으로부터 최종 깊이 지도가 아닌 텐서 $x_d \in R^{h \times w \times C_d}$를 출력한다. 한편, AdaBins 블록은 mini-ViT라는 트랜스포머(Transformer) 서브 블록과 구간 중심 추정기(Center Estimator) 서브 블록, 복합 회귀(Hybrid Regression) 서브 블록들로 구성된다. 첫 번째 mini-ViT 트랜스포머 서브 블록은 입력 영상에 대한 깊이 간격을 나누는 방법을 정의하는 구간 넓이 벡터 b(bin-width vector)와 픽셀 수준의 깊이 계산에 유용한 정보를 포함하는 크기 h×w×C의 범위-집중(Range-Attention) 지도 R을 각각 출력한다.

두 번째 구간 중심 추정기 서브 블록은 Equation 1과 같이 깊이 구간별로 구간 중심값(depth-bin-center) c(b)를 계산한다. Equation (1)에서 b는 구간 넓이 벡터를 나타낸다.

$$c(b_i) = d_{\min} + (d_{\max} - d_{\min})(b_i/2 + \sum_{j=1}^{i-1} b_j) \quad (1)$$

세 번째 복합 회귀 서브 블록에서는 Equation (2)와 같이 범위-집중(Range-Attention) 지도 R에서 구해진 각 구간별 점수인 pk와 각 깊이 구간 중심값 c(b)의 선형 조합(linear combination)으로 각 픽셀의 최종 깊이 값 $\hat{d}$를 추정한다.

$$\hat{d} = \sum_{k=1}^{N} c(b_k) p_k \quad (2)$$

3.3 경계상자와 마스크 예측

제안 모델에서 입력 영상의 관심 영역(ROI)별로 물체의 종류와 경계 상자, 그리고 마스크를 예측하는 부분은 영상

기반 물체 개체 분할(image instance segmentation)을 목적으로 개발된 본래의 Mask R-CNN 신경망과 큰 차이가 없다. 물체의 경계 상자 예측을 위한 좌표 회귀(regression)에는 기울기 폭주 현상을 방지하기 위하여, Equation (3)과 같이 소프트 L1 손실 함수(L_{bbox})가 사용된다. 물체 종류를 판별하기 위한 분류(classification)에는 교차 엔트로피 손실 함수(L_{class})를 사용한다. 또, 물체 마스크는 픽셀 단위로 분류(pixel-wise classification)가 이루어지기 때문에 Equation (3)과 같이 교차 엔트로피 손실 함수(L_{mask})가 사용된다. Equation (3)에서 y는 정답 값을, p는 예측치를, k는 관심 영역 안에 있는 픽셀의 수를 각각 나타낸다.

$$\begin{aligned} L_{bbox} &= \begin{cases} 0.5*(p-y)^2, \text{if } (p-y)^2 < 1 \\ \mid p-yvert - 0.5, else \end{cases}, \\ L_{dass} &= \sum_{c=1}^{k} y_c \log(p_c), \\ L_{mask} &= \sum_{c=1}^{k} y_c \log(p_c) \end{aligned} \quad (3)$$

3.4 NOCS 지도 예측

3차원 NOCS 표현은 동일 범주에 속한 다양한 물체들을 하나의 정규화된 3차원 좌표공간에 통합해 나타낸 것으로서, 해당 범주를 나타내는 표준화된 3차원 표현으로 해석할 수 있다. 반면에 NOCS 지도는 물체의 3차원 NOCS 표현을 카메라의 관점에서 투영해서 얻는 2차원 지도를 의미한다. 제안 모델의 확장된 Mask-RCNN 신경망 모듈은 RGB 입력 영상으로부터 물체의 종류, 경계 상자, 마스크 외에, 물체의 NOCS 지도도 예측한다. NOCS 지도 예측을 위한 회귀(regression)에는 Equation (4)와 같은 소프트 L1 손실 함수 $L(y, y^*)$를 사용한다.

$$L(y, y^*) = \frac{1}{n} \begin{cases} 5(y-y^*)^2, & \mid y-y^* \mid \le 0.1, \\ \mid y-y^* \mid -0.05, & \mid y-y^* \mid > 0.1 \end{cases} \quad (4)$$
$$\forall y \in N, y^* \in N_p,$$

Equation (4)에서 y는 정답 NOCS 지도 정답 픽셀 값을, y*은 예측된 픽셀 값을, n은 관심 영역(ROI) 내부의 마스크 픽셀 수를 각각 나타낸다.

3.5 6D 자세 예측

RGB 영상으로터 예측된 NOCS 지도는 3차원 확장 모듈에 의해 물체 마스크와 결합되어, 해당 물체의 3차원 NOCS 표현 P_n을 구하는데 이용된다. 또한 깊이 추정 모듈에 의해 예측된 깊이 지도 역시 3차원 확장 모듈에 의해 물체 마스크와 결합됨으로써, 해당 물체의 3차원 포인트 클라우드 P_m를 얻는데 이용된다. 예측이 완료된 후에는 마스크 영상을 이용하여 NOCS 지도의 물체 영역만을 잘라낸 후 컬러 코딩된 3차원 좌표를 복원함으로써 NOCS 표현 P_n을 구성한다. 자세 추론 모듈에서는 이렇게 구해진 물체의 3차원 NOCS 표현 P_n을 포인트 클라우드 P_m과 정렬(align)을 통해, 해당 물체의 크기(scale)와 회전(rotation), 변환(translation) 값을 추정한다. 이 강체 정렬 추정 문제 해결을 위해서는 Umeyama 알고리즘을 이용하고, 이상치 제거를 이해서는 RANSAC 알고리즘을 사용한다.

4. 구현 및 실험

본 논문에서는 제안 모델의 깊이 추정 성능 평가를 위해 NYU-Depth-v2 데이터 집합을 사용하였고, 자세 추정 성능 평가를 위해서는 REAL-275 데이터 집합을 사용하였다. Keras로 구현된 제안 모델은 GeForce RTX 3090 GPU가 탑재된 하드웨어와 Ubuntu 18.04.6 LTS 플랫폼에서 학습과 평가를 수행하였다.

첫 번째 실험은 제안 모델에서 채용한 AdaBins 깊이 추정기의 성능을 입증하기 위한 실험이다. 이 실험에서는 AdaBins 깊이 추정기를 대표적인 지도 학습 기반 깊이 추정기들[15-19]들, 그리고 비지도 학습 기반 깊이 추정기[22-24, 6]들과 성능을 비교하였다. 이 실험의 성능 평가 지표로는 절대 상대 오차(Absolute Relative Error, Abs Rel), 제곱근 상대 오차(Square Relative Error, Sq Rel), 평균 제곱근 오차(Root Mean Square Error, RMSE), RMSE log, 척도 불변 로그(Scale Invariant Log)와 픽셀에 대한 깊이 참값 d, 추정된 깊이 $\hat{d}$에 대하여 $\max(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i})$의 임계값을 각각 1.25, 1.25^2, 1.25^3을 적용한 정확도 지표 $\delta_1, \delta_2, \delta_3$ 등을 이용하였다. 정확도를 나타내는 $\delta_1, \delta_2, \delta_3$는 수치가 높을수록, 오차를 나타내는 나머지 평가 지표들은 수치가 낮을수록 깊이 추정 성능이 뛰어남을 의미한다.

Table 1은 벤치마크 데이터 집합인 NYU-Depth-v2를 이용해 수행한 깊이 추정 성능 실험 결과를 나타낸다. Table 1의 실험 결과에서 보듯이, 제안 모델의 AdaBins 깊이 추정기는 지도 학습 깊이 추정기들인 DAV[18], BTS[19] 보다 절대 상대 오차(Square Relative Error, Sq Rel)에서 각각 4.62%, 8.8% 더 향상된 성능을 보였고, 평균 제곱근 오차(Root Mean Square Error, RMSE)에서도 10.92%, 6.85% 더 높은 성능을 보였다. 또한, 제안 모델의 AdaBins 깊이 추정기는 비지도 학습 깊이 추정기들인 VirtualDepth[23], Auto-Recifiy Network[24]에 비해서도 임계값이 가장 엄격한 정확도 지표 δ_1에서 각각 3.2%, 10.12% 더 높은 정확도를 보였다. 이와 같은 실험 결과들을 통해, 제안 모델에서 채용한 AdaBins 깊이 추정기의 뛰어난 성능을 확인할 수 있었다.

두 번째 실험은 제안 모델의 AdaBins 깊이 추정기의 성능을 정성적으로 평가하는 실험이다. 이 실험에서는 AdaBins가 추정한 깊이 지도를 또 다른 지도 학습 깊이 추정기들인 DenseDepth[17]와 BTS[19]가 추정한 깊이 지도들과 비교해 보았다. 이 실험에서는 NYU Depth-v2 데이터 집합에서 선정한 사례들을 이용하였다. Fig. 4는 그중에서 하나의 대표 사례로서, 동일한 RGB 영상에 대해 서로 다른 깊이 추정기들이 생성한 결과인 깊이 지도들을 나타낸다.

Fig. 4에서 가장 왼쪽은 정답 깊이 지도를, 가운데는 DenseDepth와 BTS가 추정한 깊이 지도들을, 가장 오른쪽에는 제안 모델의 Adains가 추정한 깊이 지도가 위치해 있다. Fig. 4의 사례에서 볼 수 있듯이, DenseDepth와 BTS가 생성한 깊이 지도에서는 왼쪽 의자의 기하학적 정보가 일부 소실되고 오른쪽 의자의 기하학적 정보가 왜곡된 결과를 보여주는 데 반해, AdaBins가 생성한 깊이 지도에는 비교적 정보 소실과 왜곡이 적은 깊이 추정 결과를 확인할 수 있다.

Table 1. Comparison of Different Depth Estimators on the NYU-Depth-v2 Dataset

Model	δ1 (↑)	δ2(↑)	δ3(↑)	AbsRel(↓)	SqRel(↓)	RMSE(↓)	RMSElog(↓)	SILog(↓)	Log_10(↓)
Coarse/Refine network[15]	0.611	0.887	0.971	0.158	1.548	0.641	0.282	-	0.270
Residual network[16]	0.811	0.953	0.988	0.127	-	0.573	0.195	-	0.125
DenseDepth[17]	0.840	0.972	0.993	0.132	0.099	0.558	0.179	17.711	0.055
DAV[18]	0.882	0.980	0.996	0.108	-	0.412	-	-	0.10
BTS[19]	0.880	0.979	0.994	0.113	0.067	0.394	0.143	11.602	0.048
StructDepth[22]	0.817	0.955	0.988	0.140	-	0.534	-	-	0.060
virtualDepth[23]	0.875	0.976	0.994	0.108	-	0.416	-	-	0.048
Auto-Rectify Network[24]	0.820	0.956	0.989	0.138	-	0.538	-	-	0.059
Adabins[6]	0.903	0.984	0.997	0.103	0.056	0.367	0.136	10.892	0.044

Fig. 4. Depth Maps Predicted by Different Depth Estimators

세 번째 실험은 제안 모델의 물체 자세 예측 성능을 평가하기 위한 실험의 하나이다. 이 실험에서는 AdaBins로 추정한 깊이 지도(estimated depth map)를 이용한 물체 자세 예측 성능뿐만 아니라, 이것을 정답 깊이 지도(ground truth map)를 이용한 성능과도 비교해보았다. 이 실험은 물체 자세 예측을 위한 벤치마크 데이터 집합인 Real-275를 이용해 수행하였다. 성능 평가 지표로는 임계값을 50%, 25%, 10%로 적용한 3D IOU와 정답 자세를 나타내는 회전, 변환 값에 대하여 각각 m°, n cm(m={10, 15}, n={5, 10, 30})의 유사도 임계값을 적용한 정확도 지표를 이용하였다. 두 평가 지표 모두 수치가 높을수록 자세 예측 능력이 뛰어남을 의미한다.

Table 2는 이 실험의 결과표를 나타낸다. 정답 깊이 지도(GT)를 사용한 경우, 제안 모델은 3D IOU(10%) 평가 지표에 대하여 84.7%의 높은 성능을 보여주었다. AdaBins가 추정한 깊이 지도(Estimated)를 사용한 경우에도 제안 모델이 보여준 자세 예측 성능은 71.8%로서, 정답 깊이 지도를 사용한 경우에 비해 예상보다 적은 성능 차이를 보여주었다. 다만, 회전 예측 성능은 정답 깊이 지도를 사용한 경우에 비해 뚜렷한 성능 차이를 확인할 수 있었다.

네 번째 실험은 제안 모델의 물체 자세 예측 성능을 평가하기 위한 또 다른 실험으로서, 정답과 예측치의 일치 여부를 판정하는 임계값의 변화에 따른 각 물체별 변환 예측 정밀도(transla-

Table 2. Evaluation of the Proposed Model on the REAL-275 Dataset

Depth	3D IOU (50%)	3D IOU (25%)	3D IOU (10%)	10 degree, 5cm	10 degree, 10cm	15 degree, 5cm	15 degree, 10cm	15 degree, 30cm
GT	76.9	84.4	84.7	24	24.3	34.7	36.8	37.1
Estimated	33.6	55.8	71.8	7.8	8.0	12.2	13.5	15.3

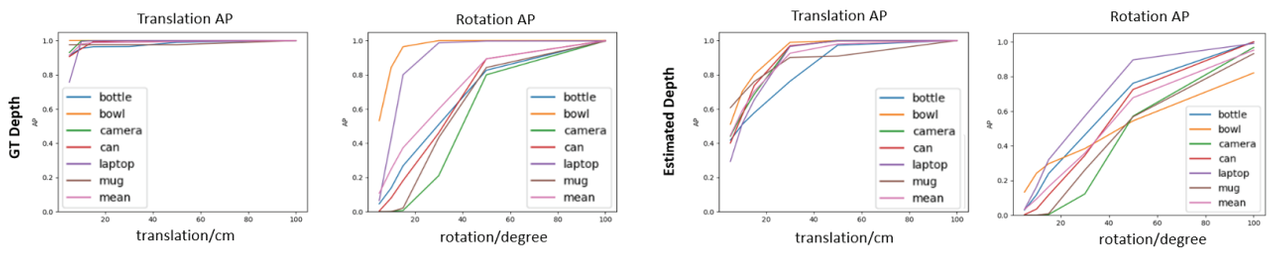


Fig. 5. Translation and Rotation AP of the Proposed Model on the REAL-275 Dataset

tion precision) 및 회전 예측 정밀도(rotation precision)를 분석하는 실험이다. 이 실험에서도 세 번째 실험과 마찬가지로, AdaBins가 추정한 깊이 지도(Estimated Depth)를 사용한 경우와 정답 깊이 지도(GT Depth)를 사용한 경우의 제안 모델의 자세 예측 결과를 서로 비교해보았다. Fig. 5는 이 실험의 결과를 보여주며, 위쪽은 정답 깊이 지도를 사용한 결과를, 아래쪽은 추정 깊이 지도를 사용한 결과를, 왼쪽은 변환 평균 정밀도(Translation AP)를, 오른쪽은 회전 평균 정밀도(Rotation AP)를 각각 나타낸다. 각 그래프의 가로축은 점차 완화되는 정답 판정 임계값을 나타내고, 세로축은 평균 정밀도(AP)를 나타낸다. 예상한 바와 같이 임계값이 완화될수록, 4개의 실험 결과 그래프에서 공통적으로 모든 물체의 변환 정밀도와 회전 정밀도는 모두 증가한 결과를 볼 수 있다. 또 Fig. 5의 위쪽 그래프들과 아래쪽 그래프들을 비교해보면, 대체로 정답 깊이 지도(GT Map)를 사용한 경우가 추정 깊이 지도(Estimated Map)를 사용한 경우에 비해 상대적으로 더 높은 변환 정밀도와 회전 정밀도를 보여주었다.

하지만 Fig. 5의 왼쪽 그래프들과 오른쪽 그래프들을 비교해보면, 임계값 완화폭이 적어지는 그래프의 왼쪽으로 갈수록 변환 정밀도에 비해 회전 정밀도의 하락 폭이 더 커지는 것을 알 수 있다. 이와 같은 실험 결과는 제안 모델이 변환 예측 능력에 비해 상대적으로 회전 예측 능력이 다소 부족함을 보여준다. 한편, 실험에 이용된 물체 중에서는 대칭성이 있는 bowl과 깊이 변동이 큰 laptop에 대한 회전 예측 성능이 다른 물체들에 비해 특히 낮은 것을 확인할 수 있었다.

다섯 번째 실험은 제안 모델의 물체 자세 예측 성능을 정성적으로 분석하기 위한 실험이다. 이 실험에서는 REAL-275 데이터 집합의 사례들을 이용하였다. Fig. 6은 제안 모델을 이용해 물체들의 6D 자세 예측을 수행한 세 가지 결과 사례들을 나타낸다. 세 가지 사례에서 보듯이, 제안 모델은 정답 깊이 지도 대신 예측된 깊이 지도를 이용하는 불리한 조건에도 불구하고, 비교적 정답에 근접한 변환 예측 결과들을 보여준다. 하지만 각 물체에 대한 회전 예측은 정답과 차이가 있음을 확인할 수 있다. 또한 첫 번째 사례와 세 번째 사례에서는 앞서 언급한 것처럼 다른 물체들보다 상대적으로 laptop과 bowl에 대한 회전 예측 정확도가 특히 낮은 것을 확인할 수 있다. 이와 같은 실험 결과를 통해, 깊이 값의 변동이 심한 물체들에 대한 변환 예측, 그리고 대칭 물체에 대한 회전 예측과 관련해서는 아직 제안 모델의 성능 개선 여지 남아있음을 동시에 확인할 수 있었다.

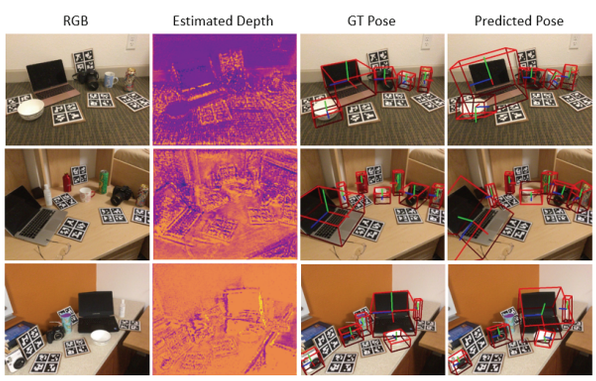


Fig. 6. Object Poses Predicted by the Proposed Model

5. 결 론

본 논문에서는 깊이 지도를 추가 입력으로 요구하는 기존 모델들과는 달리, RGB 컬러 영상만을 이용해 미지 물체들의 자세를 추정해낼 수 있는 새로운 범주-수준 자세 예측 신경망 모델을 제안하였다. 제안 모델에서는 적응형 깊이 추정기인 AdaBins를 이용하여 물체 자세 예측에 필요한 깊이 지도를 RGB 컬러 영상에서 구해낼 수 있다. 본 논문에서는 NYU- Depth-v2와 REAL-275와 같은 벤치마크 데이터 집합들을 이용한 실험을 통해, 제안 모델의 유용성과 성능을 분석하였다. 제안 모델은 별도의 3D 모델이 없어도, 또 깊이 지도가 없어도 임의 물체의 6D 자세를 추정해낼 수 있는 편리함과 높은 유용성을 제공한다. 하지만 실험적 평가에서도 언급한 바와 같이, 현재의 제안 모델은 깊이 값의 변동이 심한 물체들에 대한 변환 예측, 그리고 대칭 물체에 대한 회전 예측과 관련해서는 아직 성능 개선의 여지가 있다. 따라서 향후 연구를 통해 이러한 점들에 주목해서 현재 모델에 대한 추가 성능 개선 작업을 계속해나갈 예정이다.

References

[1] Z. He, W. Feng, X. Zhao, and Y. Lv, "6D Pose prediction of objects: Recent technologies and challenges," *Applied Science*, Vol.11, No.1, pp.228, 2021.

[2] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "LatentFusion: End-to-End differentiable reconstruction and rendering for unseen object pose prediction," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2019.

[3] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2019.

[4] M. Tian, M. H. Ang Jr, and G. H. Lee, "Shape prior deformation for categorical 6D object pose and size estimation," *Proceedings of European Conference on Computer Vision*, 2020.

[5] C. Lee, D. Shim, and H. Kim, "Deep learning based monocular depth estimation: Survey," *Journal of Positioning Navigation and Timing*, Vol.10, No.4, pp.297-305, 2021.

[6] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2021.

[7] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2018.

[8] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *Proceedings of Conference Robot Learning*, 2018.

[9] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6d pose prediction," *Proceedings of European Conference on Computer Vision*, 2018.

[10] F. Manhardt, W. Kehl, N. Navab, and F. Tombari, "Deep model-based 6D pose refinement in RGB," *Proceedings of European Conference on Computer Vision*, 2018.

[11] C. Li, J. Bai, and G. D. Hager, "A unified framework for multi-view multi-class object pose estimation," *Proceedings of European Conference on Computer Vision*, 2018.

[12] C. Wang, D. Xu, Y. Zhu, R. Martin-Martin, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2019.

[13] E. Sucar, K. Wada, and A. Davison, "NodeSLAM: Neural object descriptors for multi-view shape reconstruction," *Proceedings of 2020 International Conference on 3D Vision (3DV)*, 2020.

[14] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges, "Category level object pose estimation via neural analysis-by-synthesis," *Proceedings of European Conference on Computer Vision*, 2020.

[15] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Proceedings of NeurIPS*, 2014.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," *Proceedings of 2016 Fourth International Conference on 3D Vision (3DV)*, 2016.

[17] I. Alhashim, and P. Wonka. "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv: 1812.1194*, 2018.

[18] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkila, "Guiding monocular depth estimation using depth-attention volume." *Proceedings of European Conference on Computer Vision*, 2020.

[19] J. Lee, M. Han, D. Ko, and I. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:/1907.10326*, 2019.

[20] T. Zhou, M. Brown, N. Snavely, and D. Lowe, "Unsupervised learning of depth and ego-motion from video," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2017.

[21] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," *Proceedings of IEEE Computer Vision and Pattern Conference*, 2018.

[22] B. Li, Y. Huang, Z. Liu, D. Zou, and W. Yu, "StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation," *Proceedings of IEEE International Conference on Coputer Vision*, 2021.

[23] W. Yin, Y. Liu, and C. Shen, "Virtual Normal: Enforcing geometric constraints for accurate and robust depth prediction," *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] J. Bian, H. Zhan, N. Wang, T. Chin, C. Shen, and I. Reid, "Auto-Rectify network for unsupervised indoor depth estimation," *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

송 성 호

https://orcid.org/0000-0003-3372-4737
e-mail : ssh10032@kyonggi.ac.kr
2022년 경기대 컴퓨터공학부(학사)
2022년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 3D 비전

김 인 철

https://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(학사)
1987년 서울대학교 전산과학과(석사)
1995년 서울대학교 전산과학과(박사)
1996년 ~ 현 재 경기대학교 컴퓨터공학부 교수
관심분야 : 인공지능, 기계학습, 로봇지능