단안 깊이 추정 기술 동향

□ 김원준 / 건국대학교

요약

한 장의 이미지로부터 장면의 깊이 정보를 추정하는 기술은 자율 주행, 실내외 로봇 기반 서비스 등 다양한 응용 분야에서 널리 적용되고 있다. 심층 학습을 이용한 알고리즘이 활발히 연구되면서 이러한 단안 깊이 추정 기술의 산업 분야 적용 범위는 확대되고 있는 추세이다. 그러나, 깊이 경계 정보를 정밀하게 예측하는데 여전히 많은 어려움이 있으며, 다양한 실제 환경에서 획득한 3차원 깊이 정보 구축 또한 많은 비용이 소모되는 문제점이 있다. 본 고에서는 이러한 문제를 해결하기 위해 최근 활발히 연구되고 있는 심층신경망 기반 단안 깊이 추정 연구의 최신 동향을 소개하고자 한다. 지도 학습 기반 방법부터 최근 활발히 연구되고 있는 비지도 학습 방법까지 상세히 살펴본다. 이와 더불어 대표 방법에 대한 성능 평가 결과도 간략히 제시하고자 한다.

I. 서 론

최근 컴퓨터 비전 및 기계학습 기술의 발전으로 카메라로부터 획득한 한 장의 영상(Single Image)만을 이용하여 장면을 이해하는 기술에 대한 관심이 급증하고 있다. 특히, 장면의 깊이 지도(Depth Map)는 3차원 구조 및 객체 간 거리 등 장면 이해를 위한 주요 정보를 효과적으로 제공할 수 있어 심층신경망을 기반으로 하는 많은 방법들이 연구되고 있다. 이러한 고성능 단안 깊이 추정 기술은 자율 주행, 로봇 서비스 등에 핵심 요소로 사용되기 때문에 학계 뿐만 아니라 산업계에서 많은 관심을 가지고 연구를 진행하고 있다.

단안 깊이 추정 방법은 전통적인 영상 특징을 이용한 방법과 심층학습을 이용한 방법으로 나눌 수 있다. 먼저 영상 특징을 이용하는 방법은 주파수 계수의 통계적인 특징을 이용하여 대략적인 거리 값을 예측하는 방법[1]을 시작으로 다양한 방법에 소개되었다. 특히, 영상 분할 결과를 기반으로 각 영역별 색상, 에지, 텍스처 등의 특징을 이용하여 기하학적 구조 특성을 학습하고 깊이를 추정하는 방법이 널리 사용되었다[2]. 가장 최근에는 깊이 카메라를 이용하여 획득한 다양한 장면의 깊이 지

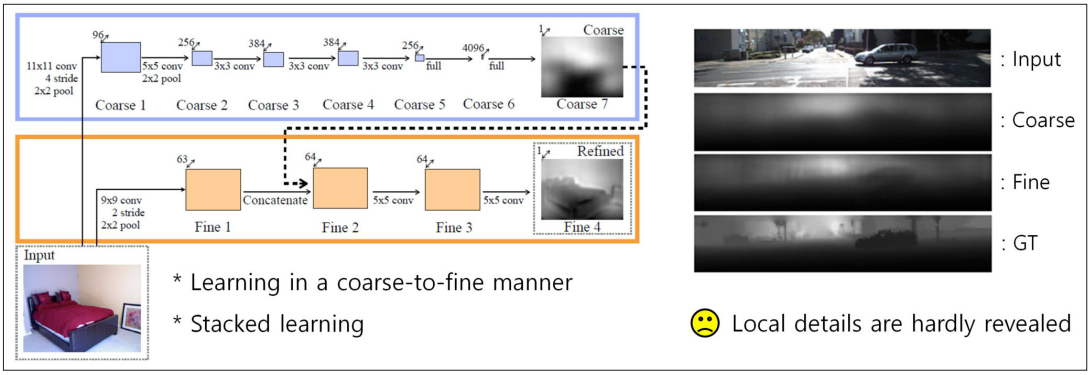
도와 현재 주어진 장면의 특징 정합을 통해 깊이 정보를 최적화하는 방법[3]도 활발히 연구되었다. 심층학습을 이용한 방법은 이와 달리, 입력 영상의 색상 값과 다양한 센서를 통해 획득한 실제 깊이 값의 관계를 학습한다. 학습이 완료된 후, 새로운 영상이 신경망에 입력되면 각 픽셀 위치의 RGB 색상 값이 상대적 깊이 값으로 변환되어 출력된다. 보다 정확한 깊이 값 예측을 위해 다양한 심층신경망 구조 및 손실 함수가 제안되어 왔으며, 최근에는 고정밀 LiDAR 센서를 이용하여 획득한 3차원 포인트 클라우드(Point Cloud) 깊이 정보 대신 단일 비디오나 스테레오(Stereo) 영상 기반 비지도(Unsupervised) 학습 기반 방법들이 활발히 소개되고 있다. 심층학습 기반 방법은 대용량 데이터를 기반으로 각 픽셀 위치에서의 색상 값과 깊이 값의 관계를 추정하기 때문에 전반적으로 전통적인 영상 특징을 이용한 방법보다 깊이 추정에 있어 우수한 성능을 보여주고 있으며, 최근에는 심층학습을 기반으로 한 방법이 대부분을 이루고 있다.

본 고에서는 단안 깊이 추정 기술 중 심층학습을 기반으로 하는 방법에 대해 심도 있게 살펴보고자 한다. II장에서는 다양한 심층신경망 구조를 기반으로 제안된 최신 기술 동향을 소개한다. III장에서는 데이터셋 및 최신 방법의 성능을 살펴보고, IV장에서 결론을 맺는다.

II. 심층신경망 기반 단안 깊이 추정 기술 동향

최근 영상 인식 분야에서 뛰어난 성능 향상을 입증한 심층학습 기술을 단안 깊이 추정에 적용하려는 시도가 늘고 있다. 기본적으로 입력 칼라 영상을 신경망에 입력으로 하여 깊이 지도(Depth Map)를 출력하는 구조를 기반으로 적용이 시작되었으며, 깊이 지도의 다양한 특성을 손실함수로 반영한 방법도 소개되고 있다. 가장 최근에는 LiDAR 센서에 의존하지 않고 스테레오 영상이나 단일 비디오 영상 내 프레임 간 관계를 바탕으로 한 비지도 학습 방식의 심층신경망 구조가 개발되고 있다. 또한, 트랜스포머를 적용한 새로운 신경망 구조를 기반으로 성능 향상을 도출하고 있다. 대부분의 심층신경망 기반의 단안 깊이 추정 방법은 깊이 경계(Depth Boundary)를 선명하게 복원하는 것을 목표로 하고 있으며, 자율 주행 플랫폼에서 동작 가능하도록 경량화 구조를 고려하기 시작하고 있다.

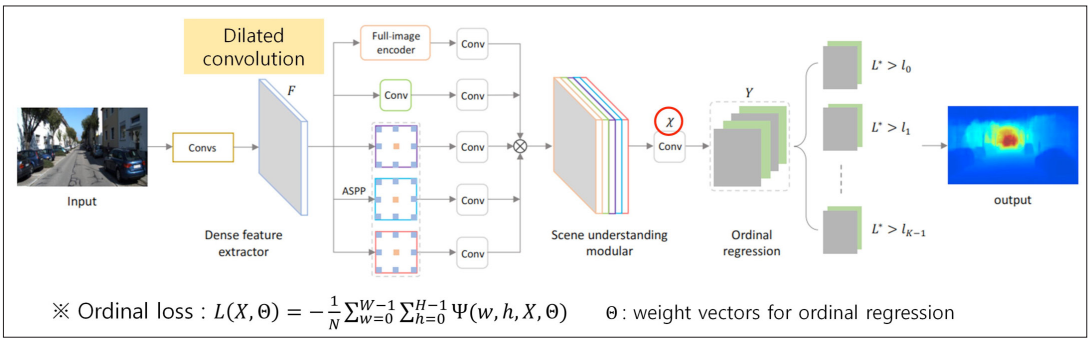
가장 먼저 영상 인식 분야에서 널리 사용되고 있는 합성곱 신경망(Convolution Neural Network, CNN) 구조가 단안 깊이 지도 생성을 위해 적용되기 시작했다. 자세히 살펴보면, Eigen[4] 등은 영상 인식에서 널리 사



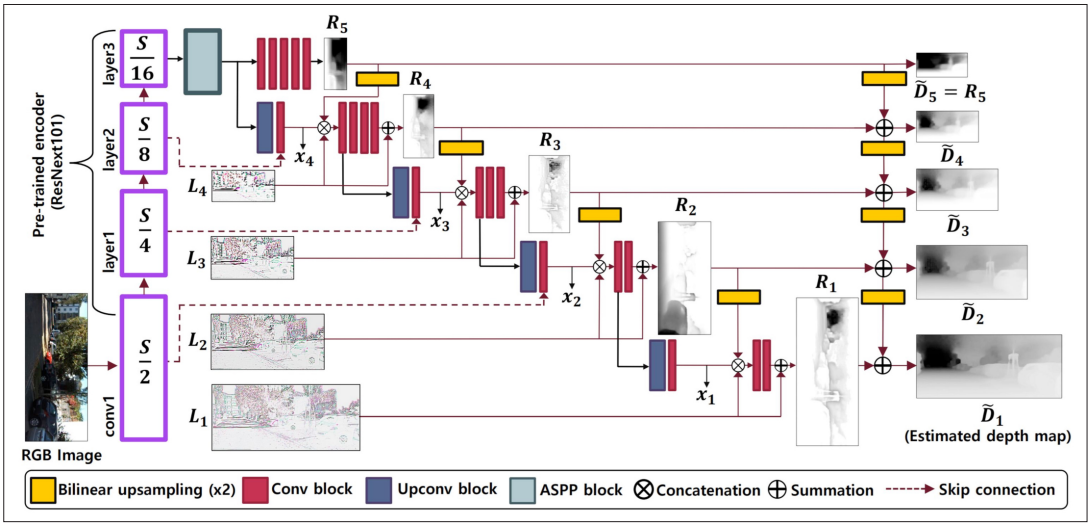
<그림 1> 합성곱 신경망 구조를 이용한 단안 깊이 추정 방법의 예[4]

용되던 적층(Stacked) 합성곱 구조를 기반으로 먼저 대략적인 깊이 지도를 생성한 후, 입력 칼라 영상과 해당 결과를 결합(Concatenation)하여 다시 지역적 상세 깊이 정보를 예측하는 2단계 구조를 제안하였다(〈그림 1〉 참조). 단순한 압축기(Encoder)-복원기(Decoder) 구조로 깊이 경계를 정확하게 예측하기 보다는 전체적인 깊이 지도 레이아웃 도출을 통해 단안 깊이 추정에 있어 합성곱 신경망의 가능성을 제시하였다. 이를 바탕으로 2015년 이후 본격적으로 심층학습 기반의 단안 깊이 추정 방법 연구가 시작되었다. Liu[5] 등은 영상 분할을 수행한 후 분할된 각 영역에서의 깊이 값을 예측할 수 있는 신경망 가지(Branch)와 분할 영역 간 깊이 유사도를 반영하기 위해 조건부 무작위장(Conditional Random Field) 기반 손실함수를 사용하는 신경망 가지를 제안하였다. Xu 등[6] 또한 조건부 무작위장을 이용하여 다중 스케일 공간에서 압축한 특징을 융합하는 방법을 통해 깊이 정보 예측 정확도를 향상 시켰다. 한편으로, Godard[7] 등은 스테레오 영상을 기반으로 시차 지도(Disparity Map)를 추정하고 이를 이용하여 스테레오 영상 간 와핑(Warping) 차이를 최소화하는 방향으로 학습을 진행하는 방식을 제안하였다. 저자는 더 나아가 동영상 프레임 간 특징 정합을 통해 카메라 매개변수를 추정하고 객체 가려짐을 효과적으로 극복하여 깊이를 예측할 수 있는 비지도 학습 방식을 제안하였다[8].

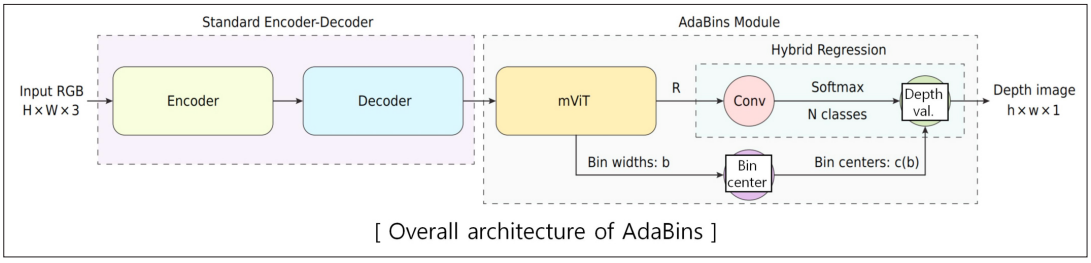
최근에는 단순히 각 픽셀 위치에서 칼라 값과 깊이 값의 관계를 회귀(Regression)를 통해 추정하는 방식에서 확장하여 깊이 정보의 속성(Attribute)을 이용하여 신경망 구조나 손실 함수를 새롭게 설계하는 방법이 소개되고 있다. Fu[9] 등은 깊이 정보는 근접 위치에서 순서(Ordinal) 관계를 형성하고 있어 각 구간에 대한 인식 문제로 손실함수를 설계하는 방법을 제안하였다. 또한, 정밀한 깊이 정보 특징 추출을 위해 다양한 크기의 수용영역(Receptive Field)를 이용한 합성곱 신경망 구조를 적용하였다(〈그림 2〉 참조). Cao[10] 등은 역시 이와 유사한 방식으로 깊이 값을 이산화(Discretization)하여 단안 깊이 추정 문제를 분류(Classification)의 문제로 재해석하였으며, 각 이산화 구간에 대한 확률 분포를 기반으로 깊이 지도를 생성하였다. 깊이 정보의 속성을 이용한 신경망 구조는 초기 방법들보다 선명하게 깊이 경계를 복원할 수 있음을 보였으나, 여전히 복잡한 배경 구조에서 성능이 저하되는 문제점이 있다. 이를 해결하기 위해 Gan[11] 등은 깊이 값의 변화가 수직 방향으로 주로 발생함을 이용하여 수직 풀링(Vertical Pooling) 모듈을 복원기에 적용하여 단안 깊이 추정 성능을 향상 시켰다. Lee[12] 등은 여러 스케일에서 깊이 정보를 적응적으로 추출하기 위해 각 스케일 공간에서 추정한



<그림 2> 순서 정보를 이용한 단안 깊이 추정 방법의 예[9]



<그림 3> 라플라시안 피라미드 기반 복원기를 이용한 단안 깊이 추정 방법의 예[13]



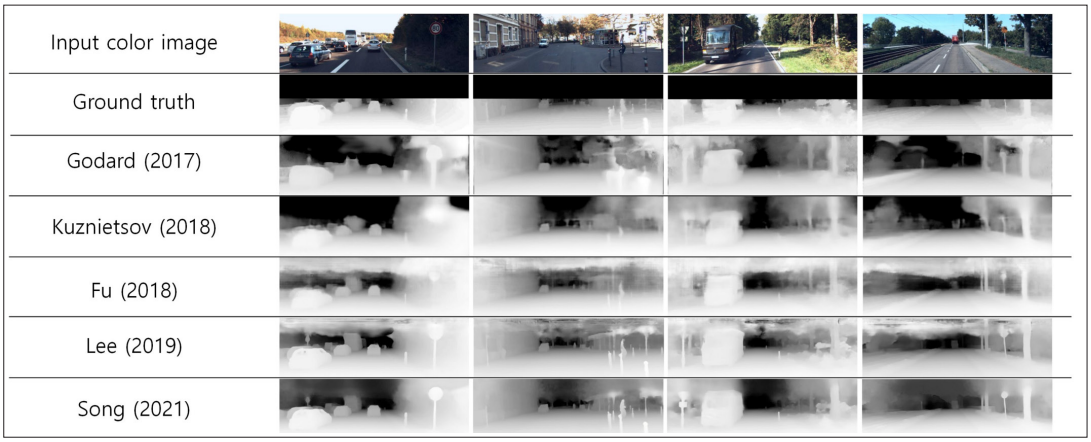
<그림 4> 트랜스포머를 적용한 단안 깊이 추정 방법의 예[16]

평면 정보(Local Planar)를 이용하여 학습을 가이드하는 새로운 방법을 제안하였다. Song[13] 등은 깊이 지도가 전역적 구조(Global Layout)와 깊이 경계로 이루어진 점에 착안하여 라플라시안 피라미드 기반 복원기를 설계하였다. 잔차(Residual) 정보를 기반으로 전역적 구조 및 지역적 상세 정보를 분리 학습하여 점진적으로 결합하는 방식을 통해 정밀하게 깊이 지도를 복원하였다(〈그림 3〉 참조). Xian[14] 등은 깊이 경계를 중심으로 다양한 픽셀 묶음을 정의하여 손실 함수를 정의하였고(Ranking Loss), 객체 영역 경계를 중심으로 같은 방식을 통해 손실함수를 정의하여 깊이 추정 성능을 향상 시켰다. 가장 최근에는 트랜스포머(Transformer)[15]를 기반으로 한 다양한 영상 인식 솔루션이 개발되고 있으며, 단안 깊이 추정을 위해 Bhat[16] 등은 기존 압축기-복원기를 통해 도출된 특징을 트랜스포머 블록을 이용하여 깊이 중앙 값을 적응적으로 추정할 수 있는 방법을 제안하였다(〈그림 4〉 참조).

최근 연구 동향을 종합해 볼 때, 데이터 측면에서는 LiDAR 센서를 통해 획득한 고정밀 3차원 포인트 클라우드 정보 대신 비지도 학습이 가능한 단일 비디오 영상을 이용하는 연구가 본격적으로 진행되기 시작했고, 알고리즘 측면에서는 트랜스포머를 이용한 전역적 특징 추출 방안 및 이를 기존 합성곱 신경망 구조와 결합할 수 있는 방법 및 학습 전략에 대한 연구가 활발히 진행되고 있다.

III. 성능 평가

단안 깊이 추정에 가장 널리 사용되는 데이터셋으로는 KITTI[17], Cityscape[18], 그리고 NYU Depth V2[19]가 있다. 먼저, 자율 주행 알고리즘 개발을 위해 대표적으로 사용되는 KITTI 데이터셋은 1,242x375 픽셀 해상도의 영상으로 구성되어 있다. KITTI 데이터셋은 총 32 장면에서 무작위로 샘플링 된 23,488장의 이미지를 학습용으로 제공하고, 나머지 29 장면에서 선택된 697장의 이미지를 성능 평가를 위해 제공한다. Cityscape 데이터 셋은 특정한 한 개의 도시에서 촬영된 KITTI 데이터 셋과 달리 20개 이상의 도시에서 촬영되어 보다 다양한 주변 환경과 해당 깊이 지도를 제공한다. 실외 환경에서 구축된 두 데이터셋과 달리, NYU Depth V2 데이터셋은 실내 환경에서 구축된 데이터셋으로 TOF(Time-of-Flight) 방식의 깊이 카메라를 이용하여 영상을 획득하였다. 따라서, 3차원 클라우드 포인트 대신 깊이 지도 이미지를 바로 제공함으로써 깊이 추



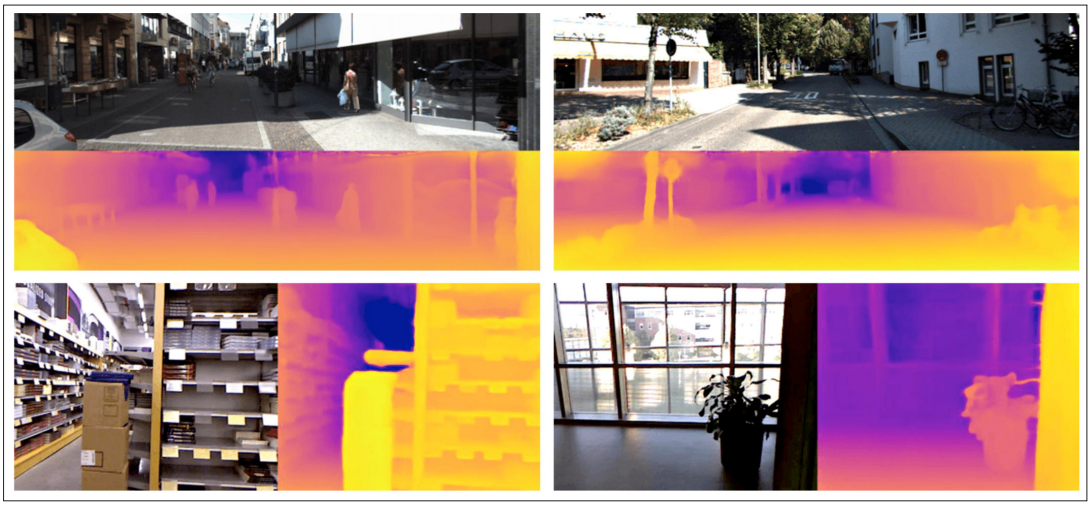
<그림 5> KITTI 데이터셋 기반 단안 깊이 추정 결과. 위쪽부터 : 입력 칼라 이미지, 정답 깊이 지도, 심층학습 기반 방법을 이용하여 예측한 결과

<그림 6> NYU Depth V2 데이터셋 기반 단안 깊이 추정 결과. 위쪽부터 : 입력 칼라 이미지, 정답 깊이 지도, 심층학습 기반 방법을 이용하여 예측한 결과

정 알고리즘 성능 평가에 널리 사용되고 있다.

먼저, 대표적인 기존 단안 깊이 추정 방법을 이용한 실험 결과를 〈그림 5〉와 〈그림 6〉에 나타내었다. 〈그림 5〉는 KITTI 데이터셋에서의 예측 결과로, 초기 방법들은 깊이 경계를 선명하게 복원하는데 어려움이 있음을 확인할 수 있다. 이와 달리, 각 스케일 특성을 반영하여 점진적으로 깊이 지도를 복원한 방법의 경우, 깊이 경계를 정밀하게 예측하고 있음을 알 수 있다. 〈그림 6〉은 NYU Depth V2 데이터셋에서의 단안 깊이 추정 결과를 보여주고 있다. 최근 심층신경망 기반의 방법은 복잡한 실내 배경 구조에서도 신뢰도 있는 깊이 추정 결과를 보여주고 있다. 〈그림 7〉은 [13] 방법의 결과를 추가적으로 보여주고 있다. 정성적 실험 결과를 통해 심층신경망 기반 단안 깊이 추정 방법이 효과적으로 산업계에 적용될 수 있음을 확인할 수 있다.

단안 깊이 추정 결과에 대한 정량적 성능 평가는 Eigen 등이 사용한 여섯 개의 평가 메트릭(Metric)이 가장 널리 사용되고 있다[4]. 각각의 메트릭은 추정한 깊이 값과 실제 깊이 값의 차이를 바탕으로 정의되어 있으며 이를 이용하여 기존 방법들에 대한 성능 평가 결과를 〈표 1〉에 나타내었다. 심층신경망 초기 모델부터, 비지도 학습 기반 방법, 깊이 속성 기반 최신 신경



<그림 7> KITTI(위쪽)와 NYU Depth V2(아래쪽) 데이터셋에서의 단안 깊이 추정 결과의 예 ([13] 방법을 이용)

<표 1> 단안 깊이 추정 방법의 정량적 성능 비교

Methods	δ1↑	δ2↑	δ3↑	REL↓	Sq REL↓	RMS↓	RMS log↓
Eigen [4]	0.702	0.898	0.967	0.203	1.548	6.307	0.282
Godard [7]	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Gan[11]	0.890	0.964	0.985	0.098	0.667	3.933	0.173
Fu [9]	0.897	0.966	0.986	0.099	0.593	3.714	0.161
Lee [12]	0.904	0.967	0.984	0.091	0.555	4.033	0.174
Song [13]	0.962	0.994	0.999	0.059	0.212	2.446	0.091
Bhat [16]	0.964	0.995	0.999	0.058	0.190	2.360	0.088

망 구조를 이용한 방법, 및 트랜스포머 기반 방법에 대한 정량적 성능을 효과적으로 비교하였으며, 최신 방법에서 신뢰도 있는 예측 성능을 확인할 수 있다. 최근 경량화 신경망 구조에 대한 연구도 활발히 진행되고 있어 임베디드 플랫폼에도 성공적으로 적용될 수 있을 것으로 예상된다.

IV. 결 론

본 고에서는 단안 깊이 추정 기술, 특히, 심층학습을 이용한 깊이 값 예측 최신 기술 동향에 대해 살펴 보았다. 컴퓨터 비전 분야에서는 영상 내 다양한 특징과 통계적 특성을 이용하여 한 장의 이미지로부터 깊이 값을 추정하는 연구가 꾸준히 진행되어 왔으나, 장면 구조를 정확히 반영하여 깊이 경계를 예측하는데 어려움이 있었다. 최근 심층학습 기술의 발전으로 설계된 특징이 아닌 데이터 학습을 통해 각 픽셀 위치에서의 색상 값과 깊이 값의 관계를 추정하는 방법이 연구되어 왔으며, 깊이 정보 속성을 반영한 다양한 신경망 구조 및 손실함수 설계를 바탕으로 괄목할 만한 성능 향상을 달성하였다. 가장 최근에는 트랜스포머를 기반으로 한 새로운 구조를 통해 실제 환경에 적용 가능한 고정밀 깊이 예측 방법들이 소개되고 있다. 이러한 연구 결과를 바탕으로 학습 파라미터 경량화 및 실내외 환경 변화에 관계 없이 강인한 동작이 가능한 심층신경망 구조가 개발된다면 인공지능 기술 기반 단안 깊이 추정 방법이 임베디드 플랫폼에도 성공적으로 적용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] A. Torralba and A. Oliva, “Depth estimation from image structure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 9, pp. 1226-1238, Sep. 2002.

[2] D. Hoiem, A. A. Efros, and M. Hebert, “Recovering surface layout from an image,” Int. J. Comput. Vis., vol. 75, no. 1, pp. 151-172, Oct. 2007.

[3] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: depth extraction from video using non-parametric sampling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 11, pp. 2144-2158, Nov. 2014.

[4] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Proc. Adv. Neural Inf. Process. Syst., Dec. 2014, pp. 2366-2374.

[5] F. Liu, C. Shen, and G. Lin, “Deep convolution neural fields for depth estimation from a single image,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 5162-5170.

[6] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, “Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 161-169.

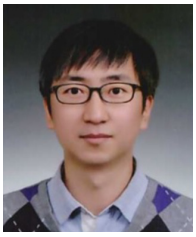
[7] C. Godard, O. M. Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2017, pp. 6602-6611.

[8] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, “Digging into self-supervised monocular depth estimation,” in Proc. IEEE Int. Conf. Comput. Vis., Nov. 2019. pp. 3827-3837.

[9] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2002-2011.

[10] Y. Cao, T. Zhao, K. Xian, C. Shen, Z. Cao, and S. Xu, "Monocular depth estimation with augmented ordinal depth relationships," IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 8, pp. 2674-2682, Aug. 2020.
[11] Y. Gan, X. Xu, W. Sun, and L. Lin, "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 232-247.
[12] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: multi-scale local planar guidance for monocular depth estimation," 2019, arXiv:1907.10326. [Online]. Available: http://arxiv.org/abs/1907.10326.
[13] M. Song, S. Lim and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 11, pp. 4381-4393, Nov. 2021.
[14] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single depth image prediction," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit, Jun. 2020, pp. 611-620.
[15] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represet., May 2021, pp. 1-12.
[16] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: depth estimation using adaptive bins," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit, Jun. 2021, pp. 4009-4018.
[17] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," Int. J. Robot. Res., vol. 32, no. 11, pp. 1231-1237, Aug. 2013.
[18] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 3213-3223.
[19] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in Proc. Eur. Conf. Comput. Vis. Oct. 2012, pp. 746-760.

필 자 소 개



김 원 준

- 2012년 8월 : 한국과학기술원(KAIST) 박사
- 2012년 9월 ~ 2016년 2월 : 삼성종합기술원 전문연구원
- 2016년 3월 ~ 2020년 2월 : 건국대학교 전기전자공학부 조교수
- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 부교수
- 주관심분야 : 컴퓨터 비전, 영상처리, 기계학습