컴퓨터 비전-자기지도학습 기반 깊이추정 방법론 동향분석

Daechan Han

***Abstract*— 자기지도 학습기반 깊이추정 방법론은 깊이 추정 센서를 대체하기 위해서 자율주행과 AR과 같은 테스크에서 매우 중요해졌다. 많은 연구자들이 자기지도 학습의 한계를 극복하기 위해서 다양한 아키택쳐를 이용한 많은 연구들이 제안 했으며, Ground truth를 직접적으로 사용하는 지도학습과 거의 유사한 성능을 달성했다. 본 레포트에서는 Self-supervised Monocular Depth Estimation의 다양한 방법론들을 비교 분석하며, 모든 분석은 많은 연구들에서 대표적으로 사용되고 있는 KITTI 데이터 셋과 더욱 어렵고 깊은 데이터를 갖고 있는 DDAD 데이터 셋에서 진행했다.**

I. 서론

딥러닝의 발달 이후 Dense Depth Map을 예측하는 Monocualr Depth Estimation은 컴퓨터 비전 분야에서 매우 중요한 부분으로 됐으며, 로보틱스와 자율주행, AR등 과 같은 다양한 어플리케이션에서 중요한 방법론으로 사용되었다. LiDAR와 stereo camera 와 같은 기존의 깊이 추정 방식들과 달리, Monocular Depth Estimation은 카메라를 한 대만 사용한다는 이점으로 가격적 부담이 적고 상용화하기 쉽다는 장점이 있다. 전통적인 방법론들은 연속적인 프레임간의 동일한 지점을 식별하기 위해서 많은 엔지니어링 기술이 필요했다. 그렇지만, 최근 Convolutional Neural Networks (CNNs)의 발달로 모든 방법론들은 CNNs 기반의 방법론으로 변화됐 으며, CNNs 기반의 방법론은 기존 전통적인 방법론들과는 달리 많은 엔지니어링 기술을 필요하지 않지만 더욱 뛰어난 성능을 보여줬다.

CNNs의 발달은 Supervised 기반의 Monocular Depth Estimation 관련 연구를 발전 시켰다. 하지만 지도 학습 방식에 필요한 Ground truth 깊이 데이터는 모으기 어려우며 학습에 필요한 데이터를 모으기에는 많은 비용이 든다는 단점이 있다. 따라서 학습에 Ground truth 깊이 데이터가 필요 없는 자기지도 학습기반 깊이추정 방법론들이 제안됐다. Garg *et al.* [7]가 처음 자기지도 학습방식이 제안했을 때는 두 카메라 간의 Disparity 관계를 사용하는 image recontruction loss

the Robotics and Computer Vision Lab, Sejong University, South Korea dchan@rcv.sejong.ac.kr

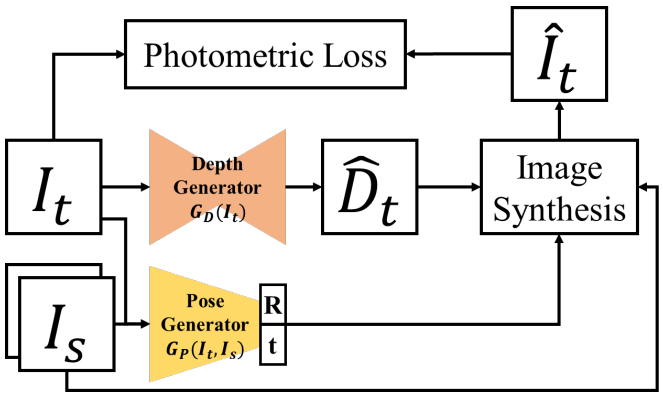


Fig. 1: **자가지도 학습 기반 깊이추정 전체 파이프라인**

를 통해서 학습을 했다. Garg *et al.*의 방법론에 영감을 받아서 Eigen *et al.* [8]는 bilinear sampling을 영상합성에 사용해서 Self-supervised monocular depth estimation의 기초를 제안했다. 그 후 카메라 두대를 사용하는 Stereo 방식이 아닌 Monocular video 만을 학습에 사용하기 위해서 Depth map과 camera motion을 동시에 예측하는 Fig. 1과 같은 방법론을 Zhou *et al.* [3]이 제안했다. 이 방법론은 Depth map과 camera motion을 예측하기 위해서 총 두 가지의 딥러닝 모델을 사용하며 예측한 두 가지 정보를 활용해 근처 프레임을 오리지널 프레임으로 새로 합성하여 비교하는 방식으로 학습한다. Zhou *et al.*가 방법론에 영감을 받아서 Eigen *et al.* [4]는 monocular video방식의 문제점이 였던 Occulusion pixel과 static pixel, dynamic objects 를 해결하기 위해서 Per-pixel minimum reprojection loss 과 Auto-Masking Stationary Pixels 방법론을 제안하였고 이후 방법론들의 기초가 되었다. Eigen *et al.* [4]이 제안한 방법론들을 바탕으로 매우 많은 연구에서 새로운 Loss 방식과 딥러닝 모델을 활용한 방법론들이 제안되었고 supervised 기반 방법론과 거의 유사한 성능을 수준까지 올라왔다.

특히 Eigen *et al.* [4]의 기본 네트워크인 ResNet [9]을 기반으로 한 Depth generator의 한계를 극복하기 위해서

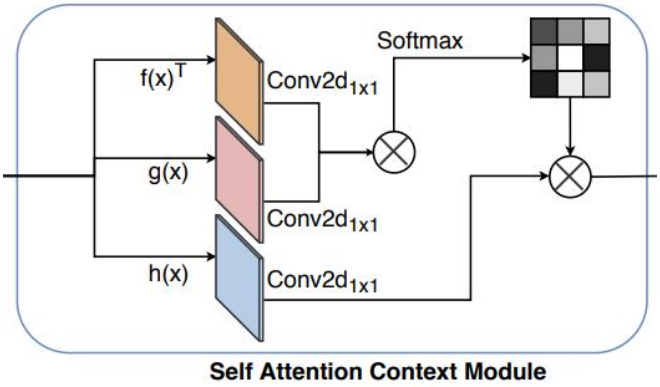


Fig. 2: **DDV [10]의 attention 방식**

Attention을 추가하거나 [10], [11], backbone network를 변경하는 [12], [2], [13], 기존 네트워크를 변경하는 [5] 등 다양한 방법론이 제안되었으며 높은 성능 향상을 이뤘다. 본 레포트에서는 Eigen *et al.* [4] 이후 딥 러닝 네트워크를 변경하여 높은 성능을 달성한 방법론들에 대해서 소개한다.

II. 방법론

2019년도에 Eigen *et al.* [4]가 Monocular video를 입력으로 하는 자기지도 학습기반 깊이추정 방법론의 기초를 다진 후 매우 다양한 연구들을 통해서 지도학습 기반 방법론들과의 성능차이가 줄어들었다. Eigen *et al.*이 매우 기본적인 딥러닝 네트워크인 ResNet [9]을 사용한 것을 문제 삼아서 여러 새로운 네트워크를 설계한 연구들이 제안되고 있으며, 제안된 방법론들에 관련해서 아래 절에서 설명하도록 한다.

A. Attention based method

DDV Self-attention은 긴 문장을 한번에 다룰 수 있다는 장점으로 Natural language processing(NLP) 분야에서 기존에 사용되던 recurrent neural networks(RNN) 방법론 [15]과 비교해서 매우 높은 성능 향상을 불러온 방법론 [16]이다. NLP 에서 성공에 영감을 받아서 Computer vision분야의 연구원들 또한 각 테스크에 적용하였고 Object Detection, Image Classification 과 같은 테스크에서 매우 좋은 성능을 보였다. 이렇게 다양한 테스크에서 높은 성능 향상을 보여준 Self-attention 방법론을 DDV [10]의 저자는 자기 지도학습 깊이 추정 방법론에 Fig. 2와 같이 적용하였고 좋은 성능을 보였다. 또한 ResNet18을 사용하던 당시 방법론들과 달리 ResNet101을 사용해서 당시 가장 좋은 깊이 추정 성능을 보였다. **MLDA-Net** MLDA-Net [11]은 Convolutional

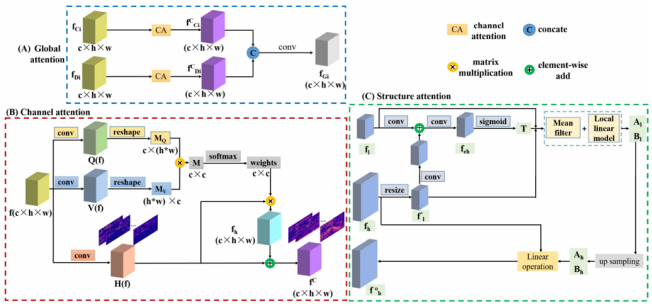


Fig. 3: **MLDA-Net [11]의 architecture** (a): color image, (b): Monodepth2 [4], (c): Monodepth2 dh [14], (d) MLDA-Net [11].

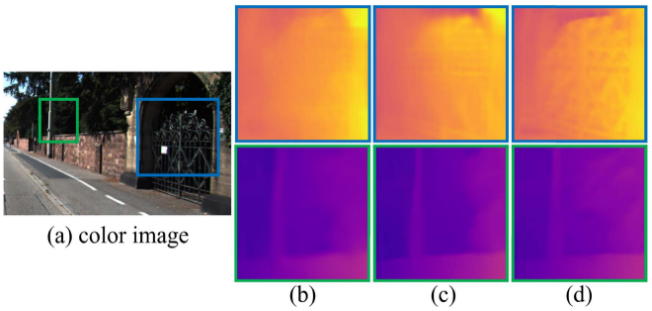
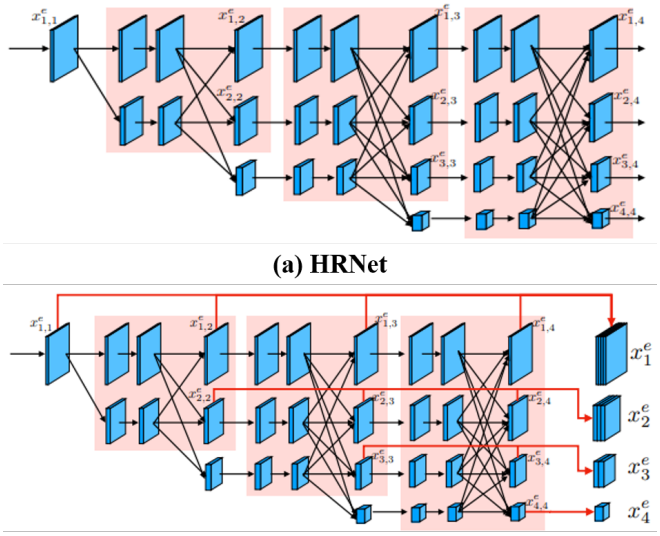


Fig. 4: **MLDA-Net [11]의 정성적 결과** (a): color image, (b): Monodepth2 [4], (c): Monodepth2 dh [14], (d) MLDA-Net [11].

Neural Networks(CNNs)의 대표적인 문제인 receptive field가 제한적이라는 문제를 해결하기 위해서 다양한 Attention 방식을 적용하였다. 먼저 Fig. 3-(A) 와 같이 CNN feature와 image feature를 Channel-Attention방식과 합께 융합하여 모든 feature의 global manner를 강화한다. 다음으로 3-(C) 와 같이 Structure Attention을 사용해서 부족했던 locality를 강화했고, locality 강화를 통한 디테일한 깊이 추정성능 향상은 Fig. 4에서 볼 수 있다. 두가지 Attention을 적용한 결과, 백본 네트워크를 ResNet18을 썼을때 기준 DDV보다 정량적으로 높은 성능과 정성적으로 디테일한 결과를 보였다.

B. Backbone Network based method

PackNet-SfM 기본적인 CNNs는 receptive field를 키우기 위해서 feature를 점진적으로 크기를 줄여가는 식으로 구성되어있다. 하지만 이렇게 구성된 네트워크의 경우 디테일한 깊이 예측이 필요한 작업에 대해서 성능이 매우 떨어지게 된다. 이때문에 전통적인 업샘플링 방법론은 정확한 깊이 예측을 복구하기 위한 디코더 계층에서 충분한 세부사항을 전파하고 보존하는 데 실패했다. 따라서 PackNet-SfM [2]에서는 중요한 공간 정보를



(a) HRNet

(b) DIFFNet

Fig. 5: **DIFFNet [12]의 백본** (a): HRNet, (b): DIFFNet [12]

보존하고 복구할수 있는 3D packing과 unpacking이라는 모듈 포함하는 PackNet을 제안했다.

Deformable-ResNet Eigen *et al.* [4]이 제안된 이후 자기지도 학습 기반 깊이 추정 방법론에는 네트워크 변경 외에도 매우 다양한 문제를 해결하기 위한 방법론들이 제안되어 왔다. 그 문제의 예시로 Occlusion 과 Static pixel, Depth Representation 등이 있다. 이렇게 다양한 문제에서 방법론들이 제안되어 왔지만 각 방법론이 다 중구난방 식으로 조합되어서 제안이 되어왔기 때문에 어떻게 조합했을때 최종적으로 좋은 성능을 보이는 지 알 수 없었다. 또한 모든 방법론들이 ResNet18만을 사용해서 성능을 보이고 다른 백본 네트워크를 사용했을 때의 성능을 리포팅 하지 않아서 각 방법론들의 잠재적인 성능 향상을 알 수 었없다. 따라서 Kim *et al.* [13]은 기존에 제안되었던 방법론들을 정리한 후 가장 좋은 조합식을 찾기 위한 실험을 했으며,네트워크에 따른 성능을 측정하기 위해서 다양한 네트워크 ResNet(18, 50, 101), Deformable-ResNet(18, 50, 101), Efficient-Net(B0,B1,B2,B4)를 사용했다. 많은 조합실험을 통해서 기존 Monodepth2의 방법론과 SoftPlus를 활용한 Depth Representation 조합이 가장 좋음을 발견했으며, CNN의 한계인 receptive field 제한을 극복하기 위해 제안된 Deformable-ResNet50 이 가장 좋은 성능을 보이는 것을 증명했다.

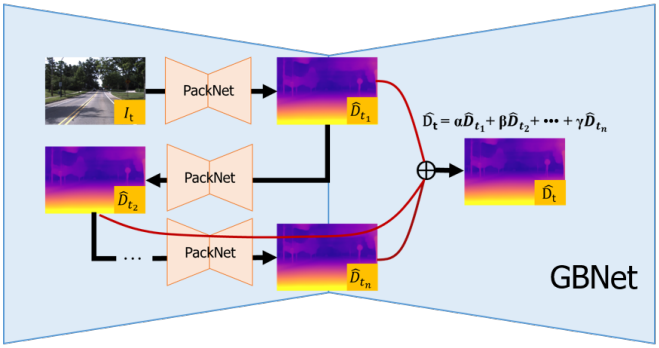


Fig. 6: **GBNet [12]의 아키텍쳐**

DIFFNet 자기지도 학습기반 깊이 추정이 자율주행의 매우 중요한 테스크가 된 가장 큰 이유는 3D 에서 보행자와 자동차와 같은 객체를 검출하기 위함이다. 따라서 예측한 깊이에서 객체를 제대로 표현하는 것은 매우 중요한 부분이며 그렇기 때문에 평가 요소에 정량적 평가뿐만이 아닌 정성적 결과를 통해서 객체를 얼마나 샤프하고 디테일하게 표현했는지가 중요한 요소로 작용한다. 그러한 이유로 Zhou *et al.* [12]는 깊이 추정에서 Semantic 정보가 중요하다고 판단을 했고, Semantic 정보를 더욱 잘 예측하기 위해서 Semantic Segmentation에서 널리 사용되고 있는 HRNet [17]을 베이스로 사용했다. Fig. 5 에서 볼 수 있는듯이 Zhou *et al.*는 단순히 HRNet을 사용하지 않고 Semantic 정보를 최대한 decoder에 전달 해줄 수 있도록 residual 한 방식으로 DIFFNet을 설계했다. 그리고 DIFFNet으로 얻은 Semantic 정보뿐만이 아니라 디테일한 정보까지 살릴 수 있도록 Attention Module 을 제안하였으며 Channel Attention과 Spatial Attention중에 Spatial Attention 만을 사용하는 것이 더욱 좋다는 것을 조합실험을 통해서 증명하였다. DIFFNet은 가장 최근 나온 방법론으로 기존 방법론들과 유사한 네트워크 크기를 가졌지만 2021년도 12월 기준 KITTI 데이터 셋에서 가장 좋은 성능을 보이고 있는 것이 특징이다.

C. ResNet based method

GBNet 기존 제안된 모든 자기지도 학습기반 깊이추정 방법론들은 전부 단일 깊이 추정 네트워크를 사용하기 때문에 오류가 있는 깊이 정보를 바로 사용해야한다는 단점이 있다. 따라서 Han [5]은 Fig. 6과 같이 딥러닝 네트워크를 계층적으로 사용하여 예측된 깊이 정보를 점진적으로 보정할 수 있는 네트워크를 설계하였다. 또한 각 네트워크에서 예측된 깊이 정보를 최종적으로 합쳐줘서 모든 예측된 결과가 상호 보와적인 관계가

될 수 있도록 구성하였다. GBNet은 비록 KITTI 데이터셋에서는 PackNet과 유사한 성능을 보여줬지만 80m까지 평가하는 KITTI 데이터셋보다 더욱 먼거리인 200m까지 평가하는 DDAD 데이터 셋에서는 높은 성능향상을 보여줬다. 먼거리 측정에서 더욱 높은 성능 향상을 이룬 것은 계층적인 네트워크를 통한 깊이 보정이 디테일한 영역에 해당하는 먼거리 픽셀에서 많은 영향을 끼친 것을 알 수 있다.

III. 결론

2019년도 이후로 매우 다양한 네트워크가 제안되었으며 모든 네트워크가 각자 다양한 문제를 해결하기 위해서 제안되어왔다는 것을 알 수 있다. 또한 각 제안된 방법론들은 기존 방법론들 보다 높은 성능을 보여줬으며 지도학습 방법론들과도 경쟁력있는 성능을 보여주는 수준까지 올라왔다는 것을 알 수 있고, 앞으로 자기지도 학습가빈 깊이추정용 딥러닝 네트워크 연구를 통해서 지도학습 기반의 방법론을 뛰어 넘을 수 있다는 가능성을 보았다.

References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. "Vision meets robotics: The kitti dataset." The International Journal of Robotics Research, 32(11):1231–1237, 2013

[2] Vitor Guizilini et al. "3d packing for self-supervised monoc u-lar depth estimation." In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2485–2494, 2020.

[3] Tinghui Zhou, Noah Snavely Matthew Brown, and David G. Lowe. "Unsupervised learning of depth and egomotion from video." In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1851–1858, 2017.

[4] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. "Digging into self-supervised monocu lar depth estimation." In Proceeding of IEEE International Conference on Computer Vision (ICCV), pages 3828–3838, 2019.

[5] Daechan Han, and Yukyung Choi. "GBNet: Gradient Boosting Network for Monocular Depth Estimation." In Proceeding of IEEE International Conference on Control, Automation and System (ICCAS), 2021.

[6] David Eigen, Christian Puhrsch, and Fergus Rob. "Depth map prediction from a single image using a multi-scale deep net work." In Proceeding of Conference on Neural Information Processing Systems (NeurIPS), 2014

[7] R. Garg, V. Kumar BG, and I. Reid. "Unsupervised CNN for single view depth estimation: Geometry to the rescue." In ECCV, 2016.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[9] Kaiming He, et al. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[10] Adrian Johnston and Gustavo Carneiro. "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume." In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020

[11] Xibin Song et al. "Mlda-net: Multi-level dual attention-based network for self-supervised monocular depth estimation." IEEE Transactions on Image Processing (TIP), 30:4691–4705, 2021.

[12] Hang Zhou and David Greenwood and Sarah Taylor. "Deep residual learning for image recognition." In Proceedings of British Machine Vision Conference (BMVC). 2021.

[13] Ue-Hwan Kim and Jong-Hwan Kim. "Revisiting selfsupervised monocular depth estimation." In Proceeding of IEEE International Conference on Computer Vision (ICCV), 2021.

[14] J. Watson, M. Firman, G. Brostow, and D. Turmukhambetov, "Selfsupervised monocular depth hints," In Proceeding of IEEE International Conference on Computer Vision (ICCV), 2019.

[15] David E Rumelhart et al. "Learning representations by back-propagating errors. Cognitive modeling," nature 5(3):1, 1988.

[16] Ashish Vaswani et al. "Attention is all you need." In Advances in neural information processing systems, pages 5998–6008, 2017.

[17] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu et al. "Deep high-resolution representation learning for visual recognition." IEEE transactions on pattern analysis and machine intelligence(TPAMI), 2020.