자기지도학습 기반 단일영상 깊이 추정 기술 동향

18011784 신정민

***Abstract*— 단일영상 기반의 깊이 추정 기술은 자율주행, SFM(Structure From Motion) 등 다양한 로보틱스 분야와 단일 영상을 활용한 증강 현실(Augmented Reality)등 다양한 영상 기술 분야에 활용할 수 있어 필요성이 점차 증가하는 기술 중 하나이다. 인공지능 기술이 점차 발전함에 따라CNN(Convolutional Neural Network)s을 활용한 지도학습 기반의 깊이 추정 방법론들이 매우 우수한 성능을 보여주며 Lidar와 같은 깊이 추정 센서를 대체할 수 있을 수준까지 올라왔다. 하지만 지도학습 기반의 방법론들은 학습을 위해 RGB 영상과 짝이 맞는 Lidar 데이터가 필수적으로 존재해야만 하며, 이러한 데이터 셋의 부재로 인해 최근에는 순수하게 영상만을 활용하는 자기지도 학습 기반의 깊이 추정 방법론들이 주요 연구 트랜드로 부상 중이다. 해당 글에서 나는 자기지도학습 기반 단일영상 깊이 추정 기술들에 대하여 전반적인 동작원리에 대해 소개할 예정이다. 또한 다양한 방법론들에 대한 실험 방법 및 결과를 비교 분석하여 깊이 추정 방법론의 현 상황을 알아보고자 한다.**

I. INTRODUCTION

깊이 정보는 예로부터 장면 이해, 자율 주행 [13]과 증강 현실 [14], [15] 등 다양한 분야에서 필수적으로 활용된다. 특히 자율주행 분야에서는 차량이 주행 경로를 판단하고자 주변 환경을 인지할 때 깊이 정보를 필수적으로 사용해야 한다. 정확한 깊이 정보를 가질수록 더 정밀한 인지 및 판단을 내릴 수 있으므로, 자율주행 시스템은 대부분 Lidar 센서를 활용하여 정밀한 깊이 정보를 추정한다. 하지만 Lidar 센서는 다른 센서들에 비교하였을 때 가격이 너무 비싸고, 휴대가 불편하며 소모성 센서라는 점에서 자율주행 차량마다 부착하기에는 무리가 있다. 이러한 문제점을 해결하고자, 카메라를 통해 깊이를 추정하는 기술들이 꾸준히 연구되는 중이다. 특히 딥러닝 기술의 발전에 따라 CNN(Convolutional Neural Networks)을 활용한 단일 영상 깊이 추정 기술 연구들 [1]–[3]이 기존의 hand-craft 방식 방법론들을 훨씬 능가하는 성능을 보여주며 빠르게 발전하는 중이다. 이러한 지도학습 기반의 깊이 추정 방법론들은 학습 때 GT로 사용할 Lidar 정보가 필수적이다. 하지만 위에서 언급한 Lidar의 여러 단점으로 인해 Lidar와 RGB 데이터가 쌍을 이루는 데이터 셋은 제한적이다. 이러한 문제를 해결하기 위해서 자기지도학습(Self-supervised Learning)기반의 깊이 추정 방법론들이 제안되어왔다. 대표적으로 Godard [4]는 네트워크를 통해 추정된 시차 영상(disparity map)을 통해 좌측 영상을 우측 영상으로, 우측 영상은 좌측 영상으로 재구축하여 비교함으로써 네트워크가 생성한 시차 영상의 좌-우가 일관성 있도록 강조하는 목적 함수를 제안하였다. 해당 방법론은 어떠한 깊이 정답 값 없이 순수하게 스테레오 영상만을 이용하여 모델을 학습하였으며 실제 동작하는 단계에서는 단일 영상을 입력으로 깊이를 추정할 수 있는 장점이 있다. 해당 방법론도 학습 때 깊이 정보를 사용하지 않으며, 평가 때는 단일 영상만으로 깊이를 추정하기 때문에 자기지도학습 기반의 단일영상 깊이 추정 기술로 볼 수는 있지만, 학습 단계에서 단일 카메라가 아닌 스테레오 카메라를 사용한다는 점에서 여전히 데이터 셋의 제약이 존재한다. 이러한 문제를 해결하고자 최종적으로 학습 때 비디오 프레임을 활용한 자기지도학습 기반 깊이 추정 기술이 단일영상 깊이 추정 분야에 한 축을 이루고 있다. 해당 글에서는 비디오 프레임을 활용한 자기지도학습 기반 단일영상 깊이 추정 기술들에 대한 동작 원리 및 기술 동향에 대해서 다루고자 한다.

II. 문제 정의 및 기술 동향

비디오 프레임을 활용한 자기지도학습 기반 단일 영상 깊이 추정 기술은 카메라의 포즈 정보와 촬영 대상의 깊이 정보를 활용한 카메라 기하학(Camera Geometry)을 통해 네트워크를 학습한다. 이 때 위의 카메라 기하학을 활용하기 위해서는 한가지 가정이 요구되는데, 이는 카메라가 항상 움직이고 있으며 촬영된 물체는 항상 정적이어야만 한다는 가정이다. 위에서도 설명했듯이 자기지도 기반의 단일영상 학습 방식은 물체의 깊이 정보 뿐만 아니라 카메라의 포즈 정보가 함께 필요하다. 그로 인해 대부분의 방법론들은 학습 단계에서 깊이를 추정하는 네트워크 뿐만 아니라 두 프레임 사이의 카메라 포즈를 추정하는 포즈 네트워크가 따로 존재한다. Zhou [5]는 깊이와 포즈 네트워크를 분리하여 학습함으로써 처음으로 자기지도 단일영상 기반의 학습 기법을 제안하였다. 해당 학습 방식의 가장 핵심적인 목적 함수는 바로 영상 재구축(Image Reconstruction) 손실 함수이다. 영상 재구축 손실 함수는 현재 프레임 I_t에서 추정한 깊이 정보와 인접

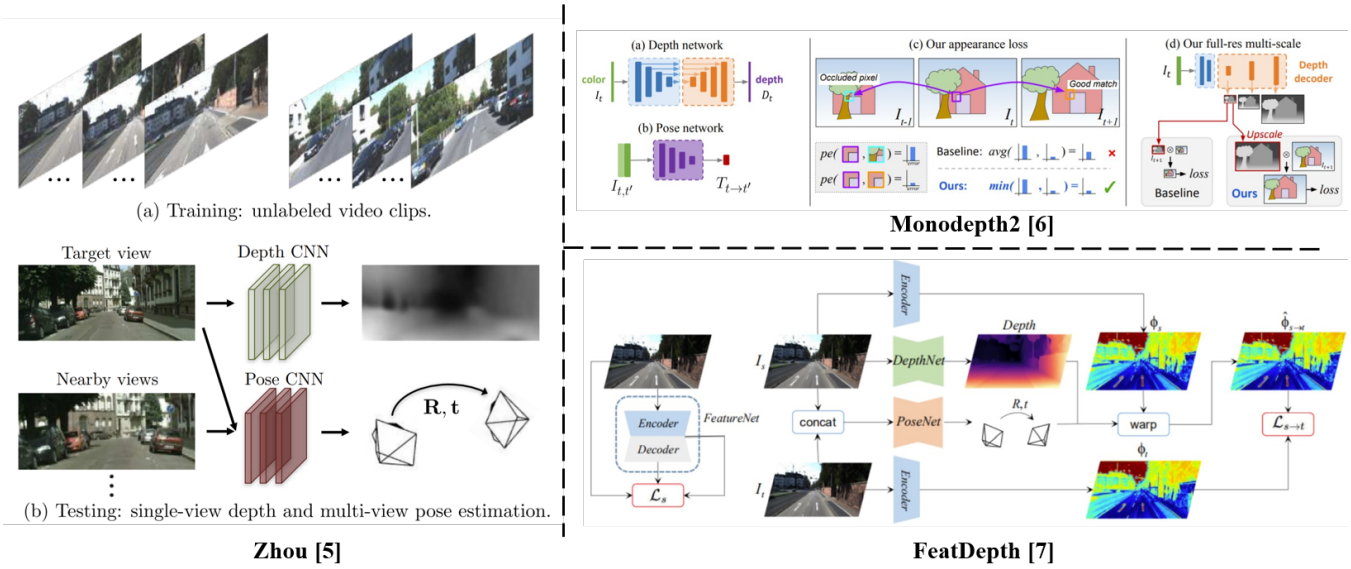

Fig. 1: **비디오 프레임을 활용한 자기지도학습 기반 깊이 추정 방법론들**

한 프레임 영상 I_{t+1}, I_{t-1} 사이의 변환행렬을 통하여 I_{t+1}, I_{t-1} 영상을 I_t로 재구축한 후 얼마나 잘 생성됐는지에 대한 에러값을 계산하여 깊이 네트워크와 포즈 네트워크를 동시에 최적화한다. 위의 과정을 통해 계산되는 영상 재구축 에러 l_p를 수식화하면 아래와 같다.

$$l_p = \sum_{t'} re(I_t, I_{(t')\to t}) \\ I_{(t')\to t} = I_{t'}\langle proj(D_t, E_{t\to t'}, K)\rangle \tag{1}$$

re는 reprojection error를 계산하는 함수를 의미하며 주로 L1 loss나 SSIM와 같은 픽셀 레벨에서의 비교 함수를 사용하며, $proj()$은 깊이 정보와 상대적인 카메라 포즈 정보를 활용해 t' 프레임의 2차원 좌표계를 t프레임으로 변환시키는 것을 의미한다. $\langle\rangle$은 sampling 연산자로 변환된 좌표계에 따라 이미지 픽셀을 보간해주는 작업을 의미한다. 위에서도 언급했듯이 자기지도 기반 단일영상 학습 방식은 한가지 기하학적 제약조건이 존재한다. 하지만 대부분의 도로주행 환경에서는 정지신호로 인해 카메라가 부착된 차량이 멈추는 경우도 있으며, 주변에 다른 차량들이 함께 움직이는 등 위의 가정이 지켜지는 경우가 매우 드물다. 이러한 예외상황을 제거하고자 Zhou [5]는 포즈 네트워크에 추가적인 디코더를 붙여 explainability mask를 추출한 후 이를 깊이 영상에 적용하여 가정에 어긋난 영역들은 학습에서 제외시켰다.

A. Monodepth2

Godard [6]는 픽셀 당 최소 재투영에러(Per-pixel minimum reprojection error), 자동마스킹(Auto-masking) 그리고 최대해상도에서의 다중 스케일 에러 계산법을 제안하면서 자기지도 기반 단일영상 학습 기법의 성능을 크게 향상시켰다. 이전의 방법론들은 I_{t-1}, I_{t+1}을 I_t로 각각 재투영시켜 계산한 재구축 에러의 평균을 계산하여 모델을 학습시켰다. 하지만 아무리 인접한 프레임이라고 할지어도 현재 프레임에서는 보이지만 이전 또는 이후 프레임에서는 보이지 않는 occlusion이 발생할 수 도 있다. 이러한 occlusion은 모델이 아무리 깊이와 포즈를 잘 추정했다 하더라도 비교 대상 자체가 존재하지 않기 때문에 loss가 크게 나와 모델의 최적화를 방해할 수 있다. 픽셀 당 최소 재투영에러는 이러한 상황을 해결하고자 이전 프레임과 현재 프레임 사이의 에러 값과 현재 프레임과 이후 프레임 사이의 에러 값을 비교하여 픽셀별로 더 작은 loss를 가지는 픽셀만을 학습에 반영하는 방식으로, 이를 통해 더 선명한 깊이 결과를 생성할 수 있게한다. 또한 자동마스킹은 현재프레임과 이전 또는 이후 프레임에 대해서 어떠한 기하학적 변환 없이 바로 reconstruction loss를 계산한 후 재투영한 프레임들로부터 계산한 loss보다 더 작은 픽셀들에 대해서 마스크처리를 하는 것을 의미한다. 이는 서로 다른 프레임끼리 비교했음에도 불구하고 동일한 위치의 픽셀들이 유사한 결과를 가지는 것이기에 카메라를 제외한 모든 물체는 정적이라는 제약조건을 위배한 픽셀로 간주하는 것이다. 이 자동마스킹은 따로 추가적인 마스크를 추출 및 연산하는 과정 없이 학습하는 단계에서 바로 사용할 수 있는 장점이 있다. 마지막으로 Godard는 기존의 깊이 네트워크에서 추출된 다중 스케일의 깊이 영상들을 각자의 해상도에서 그대로 재구축하여 에러를 계산하는 것이 아닌, 입력 해상도와 동일하게 업샘플링하

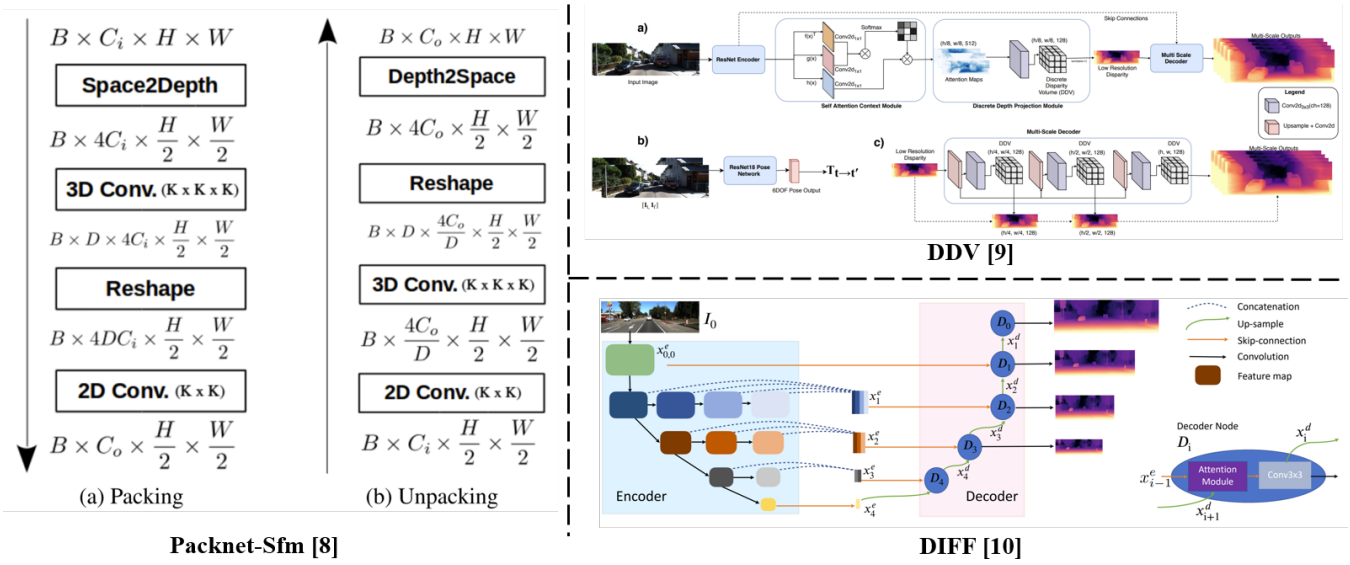

Fig. 2: **모델구조 변경을 초점으로 둔 깊이추정 방법론들**

여 재구축 에러를 계산하였다. 이러한 방식은 깊이맵의 texture-copy 현상을 제거함으로써 깊이 영상에 구멍(hole)이 생기는 현상을 방지할 수 있었다.

B. Feature-metric Loss

영상 재구축 loss는 자기지도학습 기반 깊이 추정 기술에서 매우 핵심적인 목적 함수지만, 그림과 같이 영상 속 텍스쳐가 없는 영역(textureless region)들에 대해서는 정확한 에러를 추정할 수 없는 치명적인 단점이 존재한다. 이는 정확한 포즈와 깊이를 추정하여 정밀하게 재구축을 하여서 재구축 에러가 작은 것인지, 아니면 주변에 이웃픽셀들이 실제 목표픽셀과 유사한 픽셀값을 가지고 있어서 재구축 에러가 작게 나온 것인지를 명확하게 할 수 없어 모델 학습에 큰 혼란을 주기 때문이다. 이러한 문제를 해결하고자, FeatDepth [7]는 그레이언트 값이 부각되는 feature map을 생성하는 인코더를 설계하여 feature map 추출한 뒤 이전에 추정한 깊이 맵과 포즈 정보를 통해 feature map을 재구축하여 에러를 계산하는 Feature-metric loss를 제안했으며, 이를 통해 깊이 결과의 성능을 매우 큰 폭으로 향상시켰다.

C. Packnet-Sfm

이전 서브섹션에서 설명한 Feature-metric loss와 같이 목적 함수를 잘 설계하여 깊이 추정 결과를 더 향상시켰다면 Guizilini [8]는 PackNet이라고 불리는 새로운 네트워크 설계를 통해 자기지도학습 기반의 깊이 추정 성능을 향상시켰다. PackNet은 3D컨볼루션으로 구성된 packing과 unpacking 블록을 동시에 사용함으로써 모델이 깊이 맵에 더 섬세한 디테일과 기하학적인 정보들의 표현을 극대화하도록 학습한다. 또한 부가적으로 IMU에서 추정한 프레임들 사이에 카메라 속도를 포즈 네트워크의 평행이동 정답 레이블로 사용함으로써 스케일에 강인한 깊이 추정을 가능토록 하였다.

D. Self attention and Discrete Disparity Volume

깊이추정 뿐만 아니라, semantic segmentation과 같은 pixel-level prediction 분야는 영상의 전체적인 맥락 정보(global contextual information)를 표현하는 것이 매우 중요하다. Johnston [9]은 깊이 네트워크에 self-attention 모듈을 추가함으로써 모델이 전체적인 맥락 정보를 더 잘 학습할 수 있도록 고무하였다. Self-attention 모듈은 ResNet101 Encoder를 타고 나온 가장 마지막 인코더 특징맵을 입력으로 하여 각각 Query, Key, Value로 나누어 Query와 Key를 활용해 중요도 스코어를 계산 후 이를 Value 특징맵에 곱해주는 연산을 수행한다. 또한 해당 방법론은 지도학습 기반의 깊이 추정 방법론에 영감을 얻어 깊이 맵을 이산적인 볼륨(discrete volume) 형식으로 추정한 후 채널 축으로 다 더함으로써 보다 더 선명한 깊이를 추정하였다.

E. Internal Feature Fusion

깊이 추정의 성능 향상을 위해서는 영상 속 지역적, 의미론적 표현의 손실없이 정확히 표현되어야 한다. 이러한 관점에 초점을 두어 Zhou [10]는 semantic segmentation에서 매우 좋은 성능을 보여준 HRNet을 백본으로 활용하여 깊이 성능의 정확도를 크게 향상시켰다. 더나아가 각 인코더 단계별로 추출된 특징

Method	W × H	Abs Rel ↓	Sqr Rel ↓	RMSE ↓	RMSE log ↓	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	$\delta_{1.25^3}$↑
Zhou [5]	640×192	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Monodepth2 [6]	640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
PackNet-SfM [8]	640×192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
DDV [9]	640×192	0.110	0.872	4.714	0.189	0.878	0.958	0.980
DDV(R101) [9]	640×192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
DIFF [10]	640×192	0.102	0.749	4.445	0.179	0.898	0.965	0.983
Monodepth2 [6]	1024×320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
FeatDepth [7]	1024×320	0.104	0.729	4.481	0.179	0.893	0.965	0.984
PackNet-SfM [6]	1280×384	0.107	0.802	4.538	0.186	0.889	0.962	0.981
DIFF [10]	1024×320	0.097	0.722	4.345	0.174	0.907	0.967	0.984

TABLE I: **KITTI 데이터 셋에서의 정량적 비교 결과**

맵을 융합한 뒤 채널축으로 attention 연산을 수행하여 2021년 기준 KITTI Dataset에서 최고 성능을 도달하였다. 해당 논문은 깊이 추정에서 지역 정보와 의미론적 정보들을 잘 표현하는 것이 얼마나 중요한지를 증명한다.

III. Experimental Results

A. Datasets

1) KITTI Dataset: KITTI 데이터 셋 [11]은 RGB, Grey 각각 2대의 카메라와 Velodyne Lidar, GPS/IMU 센서 시스템을 통해 다양한 자율주행 분야를 학습 및 평가할 수 있는 방대한 양의 실외 데이터 셋이다. 지도학습 기반 방법론들과 자기지도학습 기반 깊이 추정 방법론들은 대부분 실외 데이터 셋으로 KITTI 데이터 셋을 사용하고 있으며, 특히 Eigen [12]이 구분지은 학습/평가 데이터 셋을 가장 많이 활용한다. Eigen [12]이 구분한 데이터 셋은 학습데이터가 39,810 장, 평가 데이터가 697 장으로 이루어져있다.

B. Evaluation Metrics

수식 2는 단일영상 깊이 추정의 평가지표를 보여준다.

$$\textbf{Abs\_Rel}: \frac{1}{\mid D \mid}\sum_{d}^{D} \mid d^* - d \mid / d^*$$

$$\textbf{Sq\_Rel}: \frac{1}{\mid D \mid}\sum_{d}^{D} \mid\mid d^* - d \mid\mid^2 / d^*$$

$$\textbf{RMSE}: \sqrt{\frac{1}{\mid D \mid}\sum_{d}^{D} \mid\mid d^* - d \mid\mid^2}$$

$$\textbf{RMSE\_log}: \sqrt{\frac{1}{\mid D \mid}\sum_{d}^{D} \mid\mid logd^* - logd \mid\mid^2}$$

$$\delta_{\mathbf{t}}: \frac{1}{\mid D \mid} \mid \left\{ d \in D \mid max(\frac{d^*}{d}, \frac{d}{d^*}) < 1.25^t \right\} \mid \times 100\% \quad (2)$$

먼저 D는 예측한 깊이 값들의 전체 집합을 의미하며, d^*, d는 각각 정답 깊이와 예측 깊이 값을 의미한다. 수식에서도 확인할 수 있듯이 Abs_Rel와 Sq_Rel은 예측한 깊이 값과 정답 깊이 값의 차이를 계산한 후 실제 정답 값을 나누어줌으로써 계산된다. 즉 Rel 평가 지표는 동일한 오차 값을 가진다 하더라도, 실제 정답 깊이 값이 크면 분모가 함께 커지기 때문에 에러가 작게 표현되며 정답 깊이 값이 작으면 분모도 작아지기 때문에 상대적으로 큰 에러 값이 계산되기 때문에, 가까이 있는 거리 결과에 대한 정확도를 판단할 때 유용하다. RMSE와 RMSE_log는 가까운 거리에 민감한 Rel과 달리 전체 거리에 대하 일정하게 에러를 계산하는 평가 지표로 사용되며, δ는 예측한 깊이 값과 실제 깊이 값이 일정 임계치 이내로 얼만큼 속하는지를 백분율로 나타내는 평가 지표이다.

C. Quantitative Results

표 I은 KITTI데이터 셋에서 자기지도학습 기반의 깊이 추정 방법론들의 정량적 결과를 보여주고 있다. 비디오 프레임을 활용한 자기지도학습 기반의 깊이 추정을 처음 제안한 Zhou [5]와 비교하여 Monodepth2 [6]는 모든 평가지표에서 상당한 성능 향상을 보여준다. 이는 occlusion이 발생하는 상황을 해결해주는 픽셀 당 최소 재투영 에러와 기하학적 예외 상황들을 제외시키는 auto masking이 깊이 추정 학습에 매우 효과적인 것을 증빙한다. 또한 DDV [8]를 포함한 대부분의 방법론들이 동일한 방법론을 사용한다 하더라도 백본의 크기 또는 입력 해상도에 따라 전체 평가 지표에서 성능 향상이 크게 일어나는 것을 확인할 수 있으며, 이를 통해 백본의 크기와 해상도는 깊이 추정 결과의 정량적 성능에 큰 영향을 미친다는 것을 확인할 수 있다. FeatDepth [7]는 Monodepth2와 동일한 모델과 목적 함수를 사용하였으며 단지 Feature-metric loss만을 추가하였을 뿐인

데, 모든 평가 지표에서 매우 유의미한 성능 향상을 보여준다. 이는 깊이 네트워크의 주 목적함수인 영상 재구축 손실 함수가 텍스쳐가 없는 영역에 대해서는 유의미한 평가를 하지 못하기에 모델 학습을 최적화하지 못하며, Feature-metric loss가 이를 잘 보완해줄 수 있는 것을 증명한다. 마지막으로 DIFF [10]는 깊이 추정 모델이 지역적, 의미론적 정보들을 잘 표현하고 학습하는 것이 깊이 추정 정확도에 상당히 중요하다는 것을 모든 평가지표에서 최고 성능을 보여줌으로써 증명한다.

IV. Conclusion

깊이 추정 기술의 필요성은 점차 증가하는 반면에 학습을 위한 데이터 셋 취득이 매우 어려운만큼, 자기지도학습 기반 깊이 추정 연구의 중요성은 점차 증가하고 있다. 현재 비디오 프레임을 활용한 자기지도학습 기반 깊이 추정 방법론은 처음 Zhou [5]가 제안하고 나서 상당히 큰 폭의 성능 향상을 보여주며 빠르게 발전하는 모습을 보여준다. 하지만 아직까지도 지도학습 기반의 깊이 추정 방법론들과 비교하였을 때 큰 폭의 성능 격차가 존재하고 있어 실제 환경에서 사용하기에는 아직까지 무리가 있는 실정이다. 또한 깊이 추정 모델의 성능을 크게 향상시키기 위해서는 더 큰 모델의 크기와 고해상도의 깊이 영상을 생성해야만 하는데, 이는 큰 규모의 하드웨어와 연산 과정을 필요로 하기 때문에 작은 모델 크기와 수행 속도를 겸비하면서 동시에 정확한 성능 향상을 위한 깊이 추정 기술 연구가 함께 진행되어야 한다.

References

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," NIPS, 2014.

[2] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in Proceedings of the IEEE International Conference on Computer Vision, 2015.

[3] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3917–3925.

[4] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in CVPR, vol. 2, no. 6, 2017, p. 7.

[5] ZHOU, Tinghui, et al. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1851-1858.

[6] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[7] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[8] Guizilini, Vitor, et al. "3d packing for self-supervised monocular depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[9] Johnston, Adrian, and Gustavo Carneiro. "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume." Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020.

[10] Zhou, Hang, David Greenwood, and Sarah Taylor. "Self-Supervised Monocular Depth Estimation with Internal Feature Fusion." arXiv preprint arXiv:2110.09482 (2021).

[11] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

[12] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." arXiv preprint arXiv:1406.2283 (2014).

[13] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[14] Bastug, Ejder, et al. "Toward interconnected virtual reality: Opportunities, challenges, and enablers." IEEE Communications Magazine 55.6 (2017): 110-117.

[15] Ibáñez, María-Blanca, and Carlos Delgado-Kloos. "Augmented reality for STEM learning: A systematic review." Computers Education 123 (2018): 109-123.