논문 2022-17-26

비지도학습 기반의 뎁스 추정을 위한 지식 증류 기법
(Knowledge Distillation for Unsupervised Depth Estimation)

송 지 민, 이 상 준*
(Jimin Song, Sang Jun Lee)

Abstract : This paper proposes a novel approach for training an unsupervised depth estimation algorithm. The objective of unsupervised depth estimation is to estimate pixel-wise distances from camera without external supervision. While most previous works focus on model architectures, loss functions, and masking methods for considering dynamic objects, this paper focuses on the training framework to effectively use depth cue. The main loss function of unsupervised depth estimation algorithms is known as the photometric error. In this paper, we claim that direct depth cue is more effective than the photometric error. To obtain the direct depth cue, we adopt the technique of knowledge distillation which is a teacher-student learning framework. We train a teacher network based on a previous unsupervised method, and its depth predictions are utilized as pseudo labels. The pseudo labels are employed to train a student network. In experiments, our proposed algorithm shows a comparable performance with the state-of-the-art algorithm, and we demonstrate that our teacher-student framework is effective in the problem of unsupervised depth estimation.

Keywords : Unsupervised learning, Depth estimation, Knowledge distillation, Autonomous driving

Ⅰ. 서 론

최근 임베디드 하드웨어 및 빅데이터 분야의 비약적인 발달로 인하여, 다양한 실제 산업 분야에서 딥러닝 기술이 활용되는 추세이다. 딥러닝 기반의 인식기술은 미래지향적인 산업인 자율 주행, 로보틱스, 가상현실, 증강현실 등의 분야에서도 핵심 요소기술에 해당한다. 여러 가지 인식 기술들 중 영상 기반의 뎁스 추정 기술은 카메라 영상으로 화소 수준의 거리정보를 추론하는 기술로서, 3차원 인식분야에서 중요하면서도 어려운 문제에 해당한다. 단일 영상기반 뎁스 추정을 위한 딥러닝 기술은 뉴럴 네트워크 학습 시에 카메라 영상과 라이다 센서의 뎁스맵을 사용하는 지도학습과 카메라 영상만을 사용하여 학습하는 비지도학습으로 분류할 수 있다. 지도학습은 비지도학습에 비해 높은 추론 성능을 보이는 경향이 있고, 비지도학습은 학습 데이터를 낮은 비용으로 빠르게 구성할 수 있다는 장점이 있다. 본 논문에서는 비지도학습의 장점을 유지하면서 지도학습과 비지도학습의 성능 격차를 줄이고자, 지식 증류 기법 기반의 뎁스 추정 방법을 제안한다.

딥러닝 기반 뎁스 추정 알고리즘의 초기 모델은 Eigen et al.에 의하여 제안되었다 [1]. 이 초기연구에서 제안된 scale-invariant loss는 뎁스추정을 위한 연구에서 보편적으로 사용되는 추세이며, KITTI [2] 데이터셋의 뎁스 추정 벤치마크에서도 우선시되는 평가지표로 사용되고 있다. DORN [3]은 일반적으로 선형 회귀로 다뤄지는 뎁스 추정의 문제를 순서 회귀의 접근방법으로 재해석한 알고리즘이다. 이산화에 따른 에러를 낮추고자 지수적으로 간격이 증가하는 이산화 방식을 제안하고, 다수의 이진분류기를 이용하여 뎁스추정의 에러를 최소화하는 접근방법을 제안하였다. BTS [4]에서는 기존 연구에서 사용된 DenseNet [5] 기반의 인코더에서 효과적으로 뎁스 정보를 추론하기 위한 디코더 모델을 제안하였다. 이 방법의 특징은 기하학적 특성을 모델에 반영하고, 뎁스와 관련된 정보를 디코더 중간 과정에서 계산하는 local planar guidance 기법에 있다. 이와 같은 지도학습 기반의 뎁스 추정 방법은 정확도가 우수하다는 장점이 있지만, 딥러닝 모델 학습을 위한 데이터셋 구축에 카메라와 라이다의 캘리브레이션을 필요로 한다는 단점이 있다.

최근 비지도학습 기반의 뎁스 추정 기술이 활발히 진행되는 추세이다. Godard et al. [6]은 스테레오 영상에서 뎁스의 일관성을 유지하기 위한 left-right consistency loss를 제안하였다. 이후의 연구인 Monodepth2 [7]에서는 STN [8]을 활용하여 비지도학습 기반으로 뎁스추정 모델을 훈련시키기 위한 프레임워크를 제안하였으며, 정적 배경에 해당하는 화소들만을 appearance matching loss에 적용하기 위한 마스크를 제안하였다. DiPE [9]는 appearance matching loss를 보완하여 occlusion과 동적 객체로부터 발생하는 문제를 완화하였다. MT-SfMLearner [10]은 ViT [11] 기반의 지도학습 방법인 DPT [12]에 카메라의 자세를 추정하는 모듈을 추가한 비지도학습 구조를 제안하였다. 본 논문에서는 Monodepth2 [7]을 teacher network로 구성하여 pseudo

*Corresponding Author (sj.lee@jbnu.ac.kr)
Received: Jul. 4, 2022, Revised: Aug. 1, 2022, Accepted: Aug. 3, 2022.
J.M. Song: Jeonbuk National University (M.S. Student)
S.J. Lee: Jeonbuk National University (Assist. Prof.)
※ 본 연구는 삼성전자의 지원을 받아 수행된 결과임.

ⓒ IEMEK J. Embed. Sys. Appl. 2022 Aug. 17(4) 209-215
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2022.17.4.209

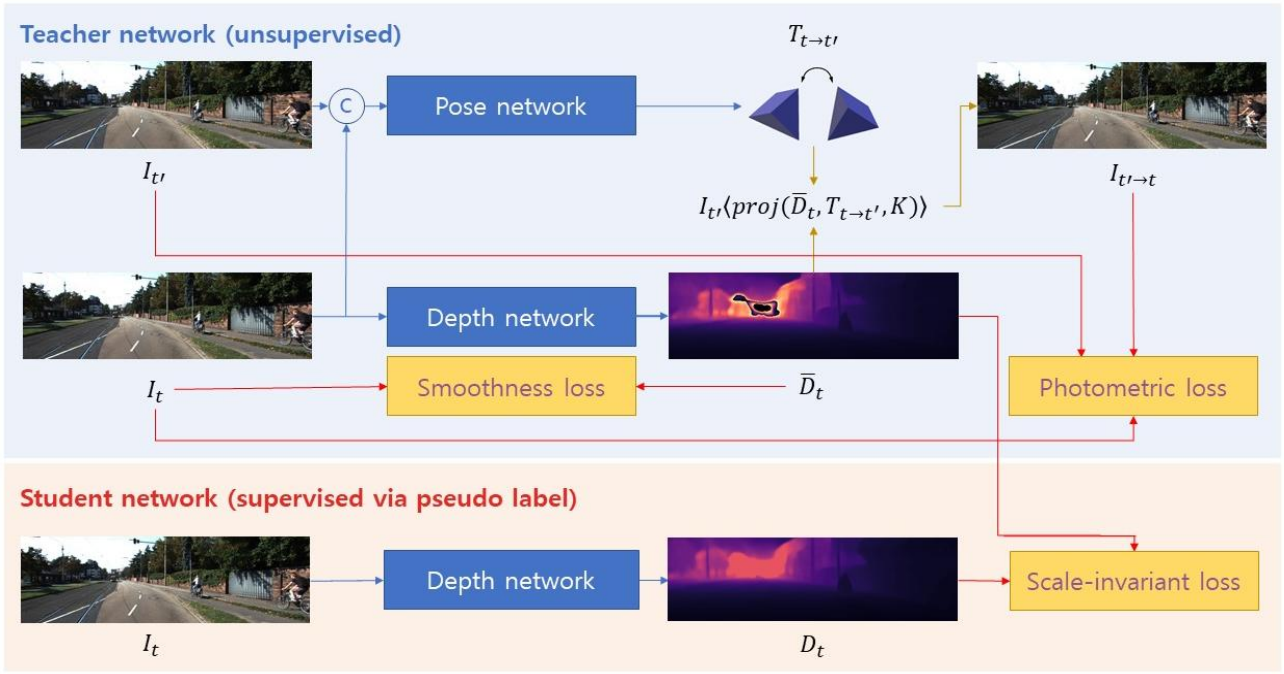

그림 1. 지식 증류 기법 기반의 뎁스 추정 모델 학습 과정
Fig. 1. Overall process for training a student depth network

depth map을 생성하고, 이를 활용하여 student network를 훈련시키는 지식증류 기법을 제안한다.

딥러닝 모델에 대한 지식 증류 기법은 Hinton et al. [13]에 의하여 제안되었다. 지식 증류 기법의 초기 목적은 스마트폰이나 사물인터넷 기기와 같은 엣지 디바이스의 한정된 연산 능력 때문에 소형화된 네트워크에서 본래의 네트워크와 비슷한 성능을 구현하는 것이었다. 지식 증류의 과정은 사전에 잘 학습된 teacher network의 추론과 ground truth를 이용하여 student network를 학습하는 과정으로 이루어진다. Teacher network의 추론 과정에서 나타나는 정보를 함께 활용하여, student network의 성능을 개선할 수 있다 [13].

지식 증류 기법은 처음 제안된 이후에 다양한 방향으로 연구되고 있으며, 최근에는 뎁스 추정을 위한 연구에도 적용되는 추세이다. Wang et al. [14]은 지식증류를 효율적이고 단계적으로 수행하기 위해서 디코더의 특징맵 사이의 pixel-wise loss를 적용하는 방법을 제안하였다. Song et al. [15]은 스테레오 영상 기반 네트워크 사이에서 단일 영상 기반 네트워크로 선택적으로 지식 증류하는 연구를 하였다. Pilzer et al. [16]은 정적인 배경과 동적인 카메라의 가정을 준수하기 위해서 복잡한 구조의 네트워크의 시차 영상에서 비교적 간단한 구조의 네트워크의 시차 영상으로 지식 증류하는 연구를 하였다. Hu et al. [17]은 빠른 추론 시간과 높은 성능을 위해서 보조 데이터셋을 활용한 지식 증류를 연구했다.

본 논문에서는 비지도학습 기반 네트워크의 추론으로 지도학습 기반 네트워크를 학습시키는 지식 증류 기법을 제안한다. Teacher network에서 추론된 뎁스맵을 pseudo label로 이용함으로써, 높은 밀도의 거리정보로 student network를 학습시킬 수 있다는 장점이 있다. 또한, 이와 같은 방법은 실제 ground truth depth 데이터를 활용하지 않기 때문에 비지도학습 기반의 방법에 해당한다. 제안하는 방법의 효과를 보이기 위하여 KITTI 데이터셋의 Eigen split [1]에서 기존 방법들과 성능을 비교하였다. 본 연구에서는 Monodepth2 [7]의 depth network와 BTS [4]의 네트워크 구조를 student network로 활용하여 실험하였으며, 제안하는 방법으로 비지도학습 기반의 최고성능 [11, 18, 19]에 근접하는 성능을 확인할 수 있었다.

Ⅱ. 본 론

1. 지식 증류 학습 체계

그림 1은 본 논문에서 제안하는 비지도 학습 기반 뎁스 추정을 위한 teacher-student 학습 프레임워크를 보여준다. 카메라 영상 I_t와 좌우 또는 전후의 인접한 카메라 영상 $I_{t'}$을 입력하여 I_t의 뎁스맵 $\overline{D}_t$를 추론한다. 그림 1의 위쪽과 아래쪽 영역은 각각 teacher network와 student network의 학습 과정을 보여준다. 그림 1의 위쪽 영역에서 ⓒ는 입력 영상을 채널 방향으로 수직하게 연결시키는 concatenation 연산을 의미하며, 빨간색, 파란색, 노란색 화살표는 각각 네트워크에서의 학습 과정, 손실함수의 연산과정, 기하학적 투영을 위한 연산 과정을 의미한다. 결정적으로 전체 네트워크 학습 프레임 워크에서 라이다 센서의 측정된 뎁스맵을 ground truth로 사용하지 않는 다는 것을 알 수 있다.

2. Teacher network 학습

Teacher network는 Monodepth2 [7]의 네트워크 구조를 사용하였다. Teacher network의 학습 과정에서 기하학적으로 재구성한 인접 영상 $I_{t\to t'}$을 계산하는 과정에서는 기존연구에서 주로 사용되는 수식 1을 사용하였다 [7]. 수식 1의 과정은 핀홀 카메라 모델 이론에 따라서 뎁스맵 $\overline{D_t}$과 카메라 내재 매개변수 K로 t'시점의 카메라 영상 $I_{t'}$을 2차원의 영상 평면에서 3차원의 월드 좌표계로 좌표계 변환을 한다. 이를 카메라 포즈 $T_{t\to t'}$로 좌표계 변환을 하면 t시점에서 t'으로 좌우 또는 전후로 시점을 전환 할 수 있다. t'시점의 3차원의 월드 좌표계 점들을 카메라 내재 매개변수 K를 활용하면 2차원의 영상 평면으로 사영할 수 있어 최종적으로 t'시점의 카메라 영상 $I_{t\to t'}$를 재구성 할 수 있다. 이 과정에서, 뎁스 추정 네트워크와 카메라 포즈를 추정하는 네트워크를 별개로 구성하였다.

$$I_{t'\to t} = I_{t'}\langle proj(\overline{D_t}, T_{t\to t'}, K)\rangle. \tag{1}$$

수식 2는 총 두 가지 항으로 구성되어 있는 데 첫 번째 항은 [20]에서 제안한 두 영상 사이의 유사도를 휘도, 대조 그리고 구조의 측면에서 계산한 SSIM으로 구성되어 있다. 두 번째 항은 L1 norm이고 이 두 가지 항을 합한 수식 2에 수식 3을 곱한 형태를 Photometric error $L_p = pe(I_a, I_b)$로 정의한다. 수식 3은 입력 카메라 영상 I_t를 재구성한 t'시점의 카메라 영상 $I_{t\to t'}$와 t'시점의 카메라 영상 $I_{t'}$에 대해서 각각 수식 2를 계산한다. 이를 비교하여 피사체가 움직이거나 카메라 자체가 움직이지 않아서 생기는 화소에 대해서 계산하지 않기 위한 마스크 μ를 생성하였다.

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - SSIM(I_a, I_b)) + (1-\alpha)\,\|I_a - I_b\|_1, \tag{2}$$

$$\mu = [\min_{t'} pe(I_t, I_{t'\to t}) < \min_{t'} pe(I_t, I_{t'})]. \tag{3}$$

Monodepth2 [7]에서 제안된 바와 같이, photometric error를 계산 하는 과정에서 수식 4와 같이 입력에 대한 전과 후의 평균이 아닌 최솟값을 이용하였다. 이를 통해서 네트워크는 인접 영상 사이에서 발생할 수 있는 폐색에 대해서 강인해질 수 있다.

$$L_p = \min_{t'} pe(I_t, I_{t'\to t}). \tag{4}$$

Direct method [21]를 따라 추정된 뎁스가 줄어드는 것을 방지하기 위해서 수식 5와 같이 평균으로 평준화된 뎁스의 역수 $d_t^* = d_t/\overline{d_t}$와 카메라 영상 I_t의 도함수로 edge-aware smoothness loss L_s를 구성하였다. 이를 통해서 네트워크는 상대적으로 어려운 피사체들의 경계부분에 대해서는 다소 낮은 가중치로 학습을 할 수 있다.

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}. \tag{5}$$

최종 손실함수는 수식 6과 같이 마스크에 해당하는 화소에 가중치를 곱한 photometric error function과 두 손실함수 사이의 가중치 λ를 곱한 edge-aware smoothness의 합으로 구성하였다. 본 논문에서는 Monodepth2 [7]과 같이 가중치 λ를 0.001로 설정하였다.

$$L = \mu L_p + \lambda L_s. \tag{6}$$

3. Student network 학습

Student network는 기본적으로 일반적인 인코더-디코더 구조의 뉴럴 네트워크를 구성하기 위해서 Monodepth2 [7]의 뎁스 네트워크의 구조를 사용하였다. 추가로 컨볼루션 커널의 채널을 늘리는 방법으로 전체 네트워크 파라미터수를 BTS [4]와 비슷한 수준으로 증가 시켰다. 그림 1에서 붉은 바탕의 과정처럼 카메라 영상 I_t을 student depth network에 입력하여 추론한 뎁스맵 D_t과 teacher depth network의 추론 뎁스맵 $\overline{D_t}$ 사이의 손실함수는 Eigen et al. [1]이 제안한 scale-invariant loss를 적용하였다.

4. 실험 구성

본 논문에서의 제안하는 방법이 기존의 다른 방법들과 비교하여 유효한지 확인하기 위해서 실험을 수행하였다. KITTI 벤치마크에서 뎁스 추정 알고리즘의 성능 평가에 사용되는 9가지 지표에 대해서 성능을 평가하고 기존 알고리즘과 비교하였다.

5. 데이터셋

Monodepth2 [7]을 따라서 Eigen split [1]을 사용하여 실험을 진행하였다. 그래서 3장의 연속된 카메라 영상이 1개인 39,810개를 training set으로 학습하고, 4,424개를 validation set으로 학습 중에 성능을 평가하였다.

Teacher network 학습에서 intrinsic parameter는 Monodepth2 [7]을 따라서 모두 동일하게 적용하여 학습하였다. Student network 학습에서는 BTS [4]와 동일하게 KITTI [2]에서 제공하는 촬영 당시의 캘리브레이션 결과를 적용하여 학습하였다.

6. 알고리즘 평가지표

뎁스 추정을 위한 많은 연구들에서 공통적으로 많이 사용되는 평가지표인 Delta threshold, Absolute relative error, Square relative error, Root mean square error, Logarithmic root mean square error, Scale-invariant error, Log10로 실험을 평가하였다. Ground truth 뎁스를 y, 네트워크에 의해서 추론된 뎁스를 y'라고 할때 Delta threshold는 $\delta = \max(\frac{y}{y'}, \frac{y'}{y})$으로 정의하고 $\delta < 1.25, \delta < 1.25^2, \delta < 1.25^3$에 대해서 만족하는 화소의 비율로 계산한다. Absolute relative error는 영상이 n개의 화소를 가질 때 $ARE = \sum_{i=0}^{n} \left|\frac{y - y'}{y}\right|$로 정의할 수 있으며 평가지표 중에서 가

표 1. KITTI Eigen split에서 기존 방법들과의 성능 비교
Table 1. Performance comparison with previous methods in KITTI Eigen split

Method	Super vision	Error metric ↓				Accuracy metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
DORN [3]	Yes	0.072	0.307	2.727	0.120	0.932	0.980	0.994
BTS [4]	Yes	0.061	0.256	2.797	0.097	0.954	0.992	0.998
DPT-Hybrid [12]	Yes	0.062	-	2.573	0.092	0.959	0.995	0.999
NeW CRFs [22]	Yes	0.052	0.155	2.129	0.079	0.974	0.997	0.999
Monodepth [6]	No	0.133	1.142	5.533	0.230	0.830	0.936	0.970
Monodepth2 [7]	No	0.107	0.817	4.788	0.198	0.871	0.956	0.979
G2S [18]	No	0.112	0.894	4.852	0.192	0.877	0.958	0.981
MT-SfMLearner [10]	No	0.112	0.838	4.771	0.188	0.879	0.960	0.982
Ours	No	0.086	0.396	3.503	0.128	0.923	0.986	0.997

표 2. Eigen split에서의 Ablation study
Table 2. Ablation study in eigen split

Method (Teacher→Student)	Super vision	Error metric ↓				Accuracy metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Monodepth2(M)	No	0.119	0.929	4.946	0.198	0.871	0.957	0.980
Monodepth2(S)	No	0.110	0.896	4.994	0.208	0.864	0.948	0.975
Monodepth2(MS)	No	0.107	0.817	4.788	0.198	0.871	0.956	0.979
Monodepth2(M)→BTS	No	0.080	0.393	3.543	0.122	0.926	0.987	0.997
Monodepth2(S)→BTS	No	0.085	0.375	3.376	0.129	0.932	0.986	0.996
Monodepth2(MS)→BTS	No	0.083	0.375	3.421	0.124	0.927	0.987	0.997

장 직관적인 지표로 사용되고 있다. 그 외에 다른 회귀문제에서도 사용되는 일반적인 평가지표로 Square relative error는 $SRE=\sum_{i=0}^{n}\frac{(y-y')^2}{y}$, Root mean square error는 $RMS=\sqrt{\sum_{i=0}^{n}(y-y')^2}$, Log10은 $\log10=\sum_{i=0}^{n}|\log_{10}y-\log_{10}y'|$, Logarithmic root mean square error는 $\log RMS=\sqrt{\sum_{i=0}^{n}(\log y-\log y')^2}$ 으로 정의된다. [1]에서 제안한 현재 뎁스 추정 문제에서 가장 중요시되 Scale-invariant error는 $SIlog=\sqrt{\sum_{i=0}^{n}(\log y-\log y')^2-(\sum_{i=0}^{n}\log y-\log y')^2}$ 으로 정의한다.

7. Eigen split 실험

표 1, 표 2에서 공통적으로 M, S, MS는 각각 Monodepth2 [7]의 단일 영상 기반 모델, 스테레오 영상 기반 모델, 단일 영상과 스테레오 영상 혼합 모델의 추론 뎁스맵을 사용했다는 것을 의미한다. 푸른 바탕의 성능지표는 높을수록, 붉은 바탕의 성능지표는 낮을수록 좋은 성능을 의미한다. Abs Rel은 Absolute relative error, Sq Rel은 Square relative error, RMSE는 Root mean square error, RMSE log는 Logarithmic root mean square error, $\delta<1.25, \delta<1.25^2, \delta<1.25^3$는 전체의 화소들 중에서 Delta 값이 threshold 값인 $1.25, 1.25^2, 1.25^3$보다 작은 화소의 비율을 의미한다.

표 1은 Eigen split [1]에서 제안하는 알고리즘의 성능을 기존의 실험들과 비교하는 자료이다. 본 논문에서의 실험 결과는 기존의 비지도학습 방법에 비하여 높은 결과를 얻을 수 있었다. Eigen split [1]에서 자주 비교되는 7가지의 지표들에서 기존의 비지도학습 방법들인 [7, 8, 11, 18]과 수치를 비교해보면 모든 지표에서 우월한 수치를 기록하며 정량적으로 우수하다는 것을 확인 할 수 있었다. 또한 비지도학습보다 지도학습에 가까운 성능을 보이며 그 격차를 많이 줄일 수 있었다는 점이 이번 실험결과에서 중요한 의미를 갖는다.

표 2는 student network는 같고 서로 다른 비지도학습 모델을 teacher network로 선정하고 학습한 결과이다. 상위 3개의 실험은 Eigen split [1]에서 비지도학습 모델 Monodepth2 [7]의 단일 영상 기반 모델, 스테레오 영상 기반 모델, 단일 영상과 스테레오 영상 혼합 모델에 대해서 학습한 실험이다. 하위 3개의 실험은 앞서 언급한 Monodepth2 [7]의 3가지 모델의 추론 뎁스맵을 pseudo label로 사용하여 student network [4]를 학습한 실험이다. 3가지 모델에 대해서 모든 지표에 대해서 성능의 향상을 확인할 수 있다. 해당 실험에서는 비슷한 성능의 서로 다른 teacher network에서 거의 유사하게 성능이 향상됨을 확인

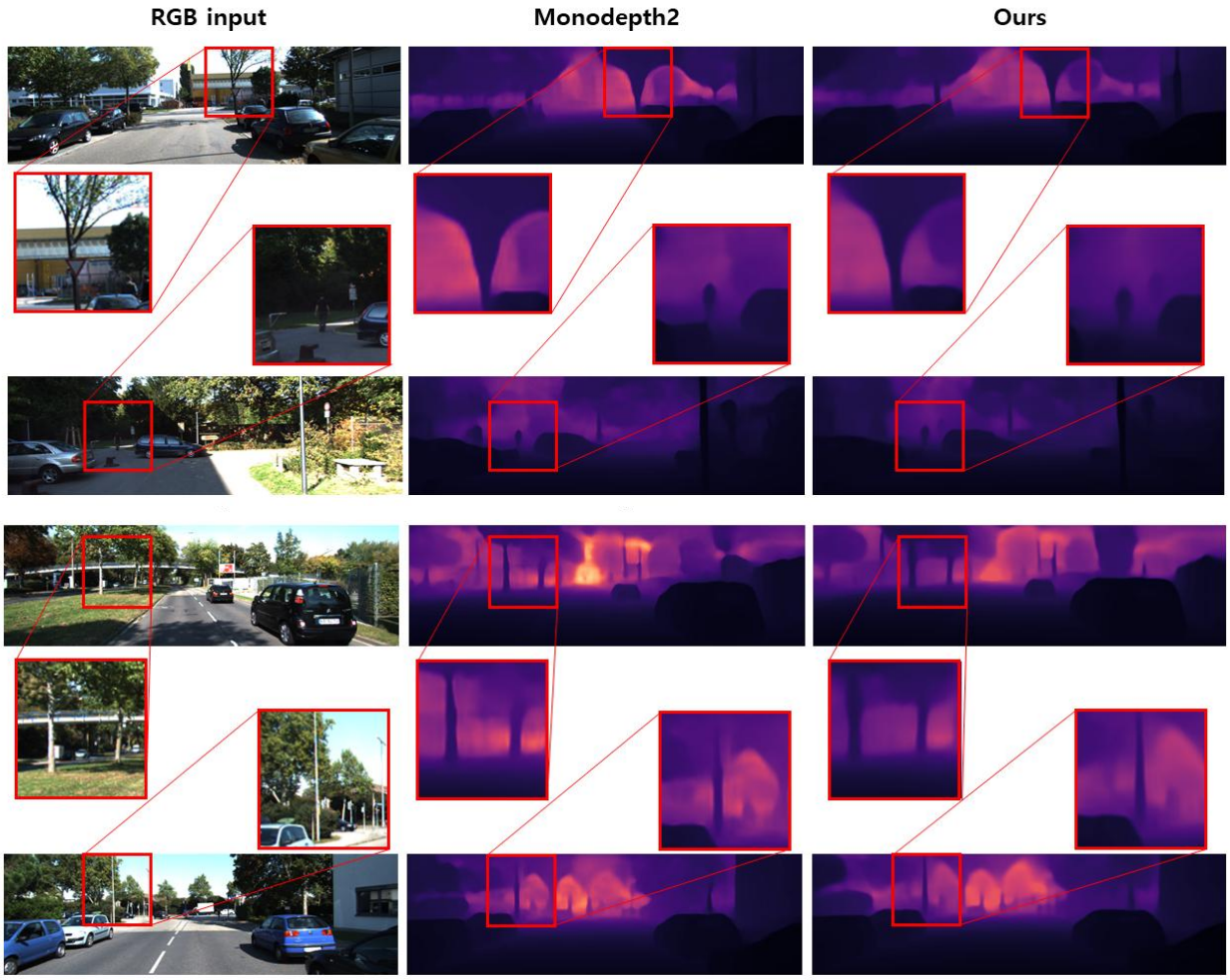

그림 2. 지식증류를 통한 뎁스맵의 선명도 개선 결과
Fig. 2. Improving object sharpness from knowledge distillation

할 수 있었다. 다시 말해서 teacher network의 성능의 유사성이 student network의 성능에서도 유사하게 반영되는 것을 확인 할 수 있었다.

그림 2는 비지도학습 방식으로 훈련한 teacher network와 본 논문에서 제안하는 지식 증류 기반으로 훈련한 student network의 추론 결과를 정성적으로 비교 및 평가하기 위한 그림을 보여준다. 예를 들어 첫 번째 영상에 대한 Student network의 추론 결과는 나무의 영역을 사실과 조금 더 가깝게 구분하였다. 두 번째 영상에서는 먼 거리의 표지판을 구분할 수 있었고 세 번째와 네 번째 영상에서는 각각 나무와 가로등의 세부 특징에 대해서 향상된 물체의 선명도를 확인할 수 있었다.

8. 실험 고찰

지식 증류가 유효할 수 있었던 근거 중의 하나는 그림 3에서 확인할 수 있다. 그림 3의 위쪽은 지도학습에 사용되는 라이다 정보로 밀도가 매우 낮다. 다소 거리 정확성이 낮은 비지도학습의 추론 결과라도 밀도가 높기 때문에 학습

그림 3. Lidar projection과 Monodepth2 추론 뎁스 맵
Fig. 3. Lidar projection and Monodepth2 depth map

에 긍정적인 영향을 준 것으로 예상된다. 그림 2에서 확인할 수 있듯이 비지도학습 기반의 teacher network는 물체의 경계에 대해서는 조금 낮은 정확도를 갖는다. Student network는 카메라 영상과 상대적으로 유사한 경계를 추론한

다. 결론적으로 높은 밀도의 pseudo label에 의한 지식 증류가 유효하다고 볼 수 있다.

Ⅲ. 결 론

본 논문에서는 비지도학습 기반의 단일 영상 기반 뎁스 추정을 위한 지식 증류 알고리즘을 제안하고 있다. 비지도학습 teacher network의 추론결과를 pseudo label로 student network를 학습하는 지식 증류 체계를 구성하였다. 결과적으로 기존의 비지도학습 기반 뎁스 추정 방법에서 주로 활용되는 photometric error를 사용하는 것 보다, depth prediction을 활용하여 직접적인 뎁스에 대한 정보가 딥러닝 모델의 성능을 높이는 데에 더 큰 효과가 있는 것을 확인할 수 있었다. 제안하는 알고리즘은 전체적으로 비지도학습이라는 점이 큰 의미가 있다. 또한 본 논문에서는 제안하는 알고리즘이 성능개선에 효과가 있음을 확인했으며, 알고리즘의 개선으로 기존의 비지도학습 방법의 SOTA와 비교할 만한 성능을 도출해낼 수 있었다.

References

[1] D. Eigen, C. Puhrsch, R. Fergus, “Depth map Prediction from a Single Image using a Multi-scale deep Network,” Advances in Neural Information Processing Systems, pp. 2366-2374, 2014.

[2] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, A. Geiger, “Sparsity Invariant cnns.” International Conference on 3D Vision, pp. 11-20, 2017.

[3] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, “Deep Ordinal Regression Network for Monocular Depth Estimation,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 2002-2011, 2018.

[4] J. H. Lee, M. K. Han, D. W. Ko, I. H. Suh, “From big to Small: Multi-scale Local Planar Guidance for Monocular Depth Estimation,” arXiv preprint arXiv:1907.10326, 2019.

[5] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, “Densely Connected Convolutional Networks,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, 2017.

[6] C. Godard, O. M. Aodha, G. J. Brostow, “Unsupervised Monocular Depth Estimation with Left-right Consistency,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 270-279, 2017.

[7] C. Godard, O. M. Aodha, M. Firman, G. J. Brostow, “Digging into Self-supervised Monocular Depth Estimation,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 3828-3838, 2019.

[8] M. Jaderberg, K. Simonyan, A. Zisserman, K. kavukcuoglu, “Spatial Transformer Networks,” Advances in Neural Information Processing Systems, pp. 2017-2025, 2015.

[9] H. Jiang, L. Ding, Z. Sun, R. Huang, “Dipe: Deeper into Photometric Errors for Unsupervised Learning of Depth and Ego-motion from Monocular Videos,” International Conference on Intelligent Robots and Systems, pp. 10061-10067, 2020.

[10] A. Varma, H. Chawla, B. Zonooz, E. Arani, “Transformers in Self-Supervised Monocular Depth Estimation with Unknown Camera Intrinsics,” arXiv preprint arXiv:2202.03131, 2022.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929, 2020.

[12] R. Ranftl, A. Bochkovskiy, V. Koltun, “Vision Transformers for Dense Prediction,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 12179-12188, 2021.

[13] Y. Wang, X. Li, M. Shi, K. Xian, Z. Cao, “Knowledge Distillation for fast and Accurate Monocular Depth Estimation on Mobile Devices,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 2457-2465, 2021.

[14] G. Hinton, O. Vinyals, J. Dean, “Distilling the Knowledge in a Neural Network,” arXiv preprint arXiv:1503.02531, 2015.

[15] K. S. Song, K. J. Yoon, “Learning Monocular Depth Estimation via Selective Distillation of Stereo Knowledge,” arXiv preprint arXiv:2205.08668, 2022.

[16] A. Pilzer, S. Lathuiliere, N. Sebe, E. Ricci, “Refine and Distill: Exploiting Cycle-inconsistency and Knowledge Distillation for Unsupervised Monocular Depth Estimation,” Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 9768-9777, 2019.

[17] J. Hu, C. Fan, H. Jiang, Xi. Guo, Y. Gao, X. Lu, T. L. Lam, “Boosting Light-weight Depth Estimation via Knowledge Distillation,” arXiv preprint arXiv:2105.06143, 2021.

[18] H. Chawla, A. Varma, E. Arani, B. Zonooz, “Multimodal Scale Consistency and Awareness for Monocular Self-supervised Depth Estimation,” International Conference on Robotics and Automation, pp. 5140-5146, 2021.

[19] J. Yan, H. Zhao, P. Bu, Y. S. Jin, “Channel-Wise

Attention-Based Network for Self-Supervised Monocular Depth Estimation," International conference on 3D Vision, pp. 464-473, 2021.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image Quality Assessment: from Error Visibility to Structural Similarity," Transactions on Image Processing, Vol. 13, No. 4, pp. 600-612, 2004.

[21] C. Wang, J. M. Buenaposada, R. Zhu, S. Lucey, "Learning Depth from Monocular Videos using Direct Methods." Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 2022-2030, 2018.

[22] W. Yuan, X. Gu, Z. Dai, S. Zhu, P. Tan, "New Crfs: Neural Window Fully-connected Crfs for Monocular Depth Estimation," arXiv preprint arXiv:2203.01502, 2022.

Jimin Song (송 지 민)

2022 Department of Electronic Engineering from Jeonbuk National University (B.S.)

2022~Department of Electronic Engineering from Jeonbuk National University (M.S.)

Field of Interests: : Artificial intelligence, Computer vision, Deep learning, Robotics

Email: jimin_song@jbnu.ac.kr

Sang Jun Lee (이 상 준)

2011 Electrical Engineering from POSTECH (B.S.)

2018 Electrical Engineering from POSTECH (Ph.D.)

Career:

2018~2020 Samsung Advanced Institute of Technology (Senior Researcher)

2020~Jeonbuk National University (Assist Prof.)

Field of Interesst: : Artificial intelligence, Computer vision, Deep learning, Robotics

Email: sj.lee@jbnu.ac.kr