Revisiting Self Supervised Depth Estimation

Wednesday, 23. June 2021 10:09 PM

Challenges :

- Lidar sensors have depth but are expensive => cameras are useful
- Supervised Depth is expensive

=> Self-Supervised methods :

- require geometric constraints to enforce consistency :

=> assumes scenes are static and Lamebertian (ideal reflecting surface)

- Learning optical flow as a side task enhances the performance (group smoothen loss are relevant here for learning dynamic object masks)

Uncertainty Estimation

- Can avoid overconfident predictions
- Aids decision making process
- Uncertainty cam arose from both model and data
- Methods : Monte Carlo Dropout, Laplace Approximation

ImageNet features get streamed efficiently to other computer vision tasks

Depth Representations

1) Disparity : the inverse of depth = 1/x
   Can represent distant objects stabily, suffers when adjacent objects appear in sc
2) Scaled Disparity
   Scales the disparity to a pre defined range for stability
   => extra hyperparameters

$$x' = \sigma_{\text{min}} + (\sigma_{\text{max}} - \sigma_{\text{min}}) \cdot x, \tag{2}$$

where σ_{max} and σ_{min} are the maximum and minimum values disparity can take. Finally, the depth values are $d = 1/x'$.

ess loss and L1/2 sparsity

ence (d <<1)

3) Softplus
   Directly predict the depth values => no hyperparameters

Softplus [32] representation lets neural networks directly predict depth values rather than disparities as follows:

$$d = \log(\exp(x) + 1). \tag{3}$$

The softplus representation avoids the case where the predicted depth values equal to zeros. This setting eliminates the need for setting hyper-parameters such as minimum disparity and lets CNNs learn the optimal values from data.

<u>Illumination Variation</u>

Not all scenes are Lamertian and therefor the supervision needs correction :

1) Brigthness Transformation => models change of intensity with affine transforma
   and depth estimation
2) Structural Similarity : computes the structural similarity between patches
   De facto approach

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\mu_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)}, \tag{5}$$

3) Depth error weighted
   Penalizes patches where the depth estimation is not consistent

<u>Occlusion Handling</u>

Hinder reconstruction, induce high photometric penalty

1) **<u>Minimum Projection</u>**
   <u>Handles occlusions by taking a minimum operation rather than averaging, han</u>

2) **Depth Consistency**

tion, improves the motion

dles out of pixel values

2) Depth Consistency
Uses the fact that the depth values become multi value when occlusions occur
Project depth at a pixel position -> obtain perspective point -> apply camera m
only counts pixels where transformed source is smaller than target source -> n
source and target switching)

Dynamic Objects
Violate static scene assumption -> gradients deterioate

1) Auto Masking
Removes scenes where camera does not move and the objects that move with
camera (appear as holes with infinite depth)

2) Uncertainity Modeling
Also used for the Lambertian assumption, it can handle moving objects

rary approaches commonly employ the heteroscedastic aleatoric uncertainty [21]—regarding dynamic objects as observation noise—as follows:

$$L = \frac{\min_{t'} pe(\boldsymbol{I}_{t' \to t}, \boldsymbol{I}_t)}{\sum_t} + \log \sum_t, \tag{9}$$

3) Motion Map (learns 3D motion map)
Can account for all types of motions with rigid translations
Key is L1/2 sparsity loss :

$$L_{1/2} = 2 \sum_{i \in \{x,y,z\}} \langle |T_i| \rangle \iint \sqrt{1 + |T_i| / \langle |T_i| \rangle} dudv, \tag{10}$$

where $\langle |T_i| \rangle$ is the spatial average of T_i. In addition, the motion map approach should be applied after a few training epochs and fed with estimated depth maps for stable learning.

<u>otion -> compare depth</u>
<u>ot symmetric (requires</u>

<u>the same velocity as the</u>

CNN Architectures
Compare ResNET, Deformable ResNET and EfficientNet

For Softplus and Dispatriy : inverse depth maps
Just L1/2 loss is used and smootheness loss, cyclic loss can corrupt the gradients
Tradeoff between representation and learning stability (Softplus diverges often)
Combinations of different representations are not trivially good or bad
Auto Mask and Motion Map whenever possible
DeResNet (50) performs best
Scaled Disparity representations make networks learn smaller depth values, MR loss
depth consistency makes networks learn much smaller depth

Table 2. Inter-dependency between various learning approaches. The double-edged line in the middle separates the first and secon
The overall leading performance metrics are in bold while the leading performance metrics in each group are underlined. Repr, D
AM, Uncrt and MM in the table stand for representation, depth-error weighted SSIM, occlusions, auto-masking, uncertainty mode
motion map, respectively. †-indicates the previous state-of-the-art configuration [11]

ID	Repr	Illumination		Occ	Dynamic Object			ARD	SRD	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<$
		$aI+b$	DW		AM	Uncrt	MM							
R0	$1/x$	-	-	-	-	-	-	0.123	1.188	5.148	0.202	0.867	0.954	0
R1	$1/x$	-	-	MR	-	-	-	0.121	1.044	5.074	0.198	0.869	0.957	0
R2	$1/x$	-	-	MR	✓	-	-	0.119	0.896	4.882	0.196	0.870	0.958	0
R3	$1/x$	✓	-	MR	-	-	-	0.122	1.050	5.083	0.199	0.869	0.957	0
R4	$1/x$	-	✓	MR	✓	-	-	0.118	0.872	4.805	0.196	0.872	0.958	0
R5	$1/x$	✓	✓	MR	-	-	-	0.120	1.012	5.027	0.197	0.872	0.958	0
S0	(0.1,100)	-	-	-	-	-	-	0.122	1.095	5.124	0.202	0.868	0.954	0
S1	(0.1,100)	-	-	MR	-	-	-	0.121	1.052	5.071	0.198	0.871	0.957	0
S2†	(0.1,100)	-	-	MR	✓	-	-	0.117	0.899	4.882	0.196	0.872	0.958	0
S3	(0.1,100)	✓	-	MR	-	-	-	0.121	1.014	5.044	0.198	0.867	0.957	0
S4	(0.1,100)	-	✓	MR	✓	-	-	0.121	0.938	4.933	0.199	0.868	0.956	0
S5	(0.1,100)	✓	✓	MR	-	-	-	0.118	1.002	5.055	0.197	0.873	0.957	0
L0	log	-	-	-	-	-	-	0.131	1.446	5.403	0.211	0.963	0.951	0
L1	log	-	-	MR	-	-	-	0.120	0.959	4.971	0.197	0.871	0.958	0
L2	log	-	-	MR	✓	-	-	0.116	0.866	4.884	0.196	0.874	0.958	0
L3	log	✓	-	MR	-	-	-	0.120	1.027	5.040	0.197	0.871	0.957	0
L4	log	-	✓	MR	✓	-	-	0.121	0.954	4.953	0.199	0.866	0.957	0
L5	log	✓	✓	MR	-	-	-	0.120	1.001	5.034	0.197	0.872	0.957	0
M0	log	-	-	MR	-	-	-	0.120	0.959	4.971	0.197	0.871	0.958	0
M1	log	-	-	MR	✓	-	-	0.116	0.866	4.884	0.196	0.874	0.958	0
M2	log	-	-	MR	-	✓	-	0.125	0.940	5.010	0.198	0.853	0.954	0
M3	log	-	-	MR	✓	-	✓	**0.114**	**0.825**	**4.706**	**0.191**	**0.877**	**0.960**	0
M4	log	-	-	MR	-	✓	✓	0.126	0.857	4.954	0.197	0.843	0.953	**0**
D0	log	-	-	DC	-	-	-	0.124	1.046	5.069	0.203	0.864	0.953	0
D1	log	-	-	DC	✓	-	-	0.119	0.919	4.925	0.198	0.867	0.955	0
D2	log	-	-	DC	-	✓	-	0.129	0.924	5.068	0.203	0.840	0.949	0
D3	log	-	-	DC	✓	-	✓	0.117	0.845	4.918	0.199	0.865	0.954	0
D4	log	-	-	DC	-	✓	✓	0.134	0.971	5.870	0.216	0.811	0.936	0
C0	log	-	-	M+D	-	-	-	0.118	0.964	4.960	0.195	0.871	0.958	0
C1	log	-	-	M+D	✓	-	-	0.118	0.878	4.842	0.196	0.867	0.957	0
C2	log	-	-	M+D	-	✓	-	0.127	0.993	5.024	0.198	0.854	0.954	0
C3	log	-	-	M+D	✓	-	✓	0.116	0.878	4.931	0.195	0.869	0.957	0
C4	log	-	-	M+D	-	✓	✓	0.136	1.069	6.332	0.224	0.811	0.931	0

. vaires the depth range,

d stages.
W, Occ,
eling and

1.25^3
.978
.980
.981
.980
.980
.980
.978
.980
.980
.980
.980
.980
.976
.980
.981
.980
.979
.980
.980
.981
.982
.982
.983
.978
.980
.981
.980
.978
.981
.981
.981
.981
.974

Table 3. The effect of CNN architectures on the monocular depth estimation performance. A pertinent
enhances the performance (DeResNet-50). All models display satisfactory processing speeds.

Architecture	Epochs	ARD	SRD	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$
ResNet-18	< 20	0.114	0.825	4.706	0.191	0.877	0.960
ResNet-50	< 20	0.110	**0.735**	4.606	**0.187**	0.880	**0.961**
ResNet-101	< 20	0.112	0.756	4.655	0.191	0.875	0.960
DeResNet-18	< 20	0.130	0.907	5.014	0.208	0.845	0.948
DeResNet-50	< 20	**0.108**	0.737	**4.562**	**0.187**	**0.883**	**0.961**
DeResNet-101	< 20	0.114	0.832	4.752	0.195	0.876	0.957
EfficientNet-B0	< 5	0.120	0.804	5.025	0.195	0.852	0.956
EfficientNet-B1	< 5	0.140	0.948	5.541	0.213	0.811	0.943
EfficientNet-B2	< 5	0.124	0.860	4.732	0.190	0.859	0.960
EfficientNet-B4	< 10	0.113	0.864	4.785	0.189	0.875	0.960

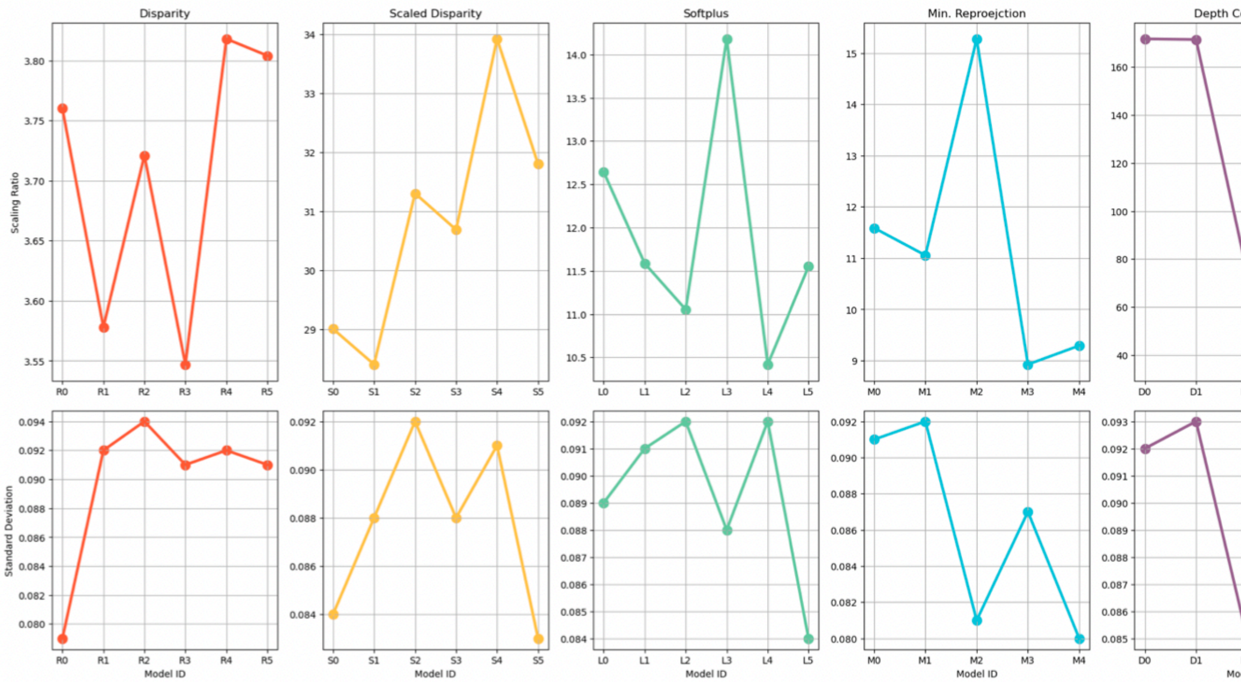
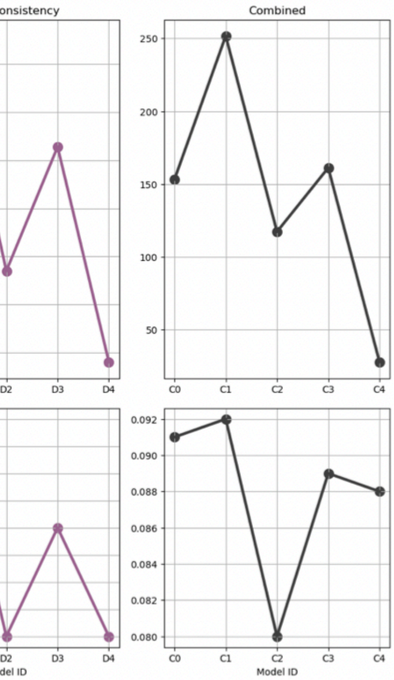


Figure 1. The scaling ratio and standard deviation of depth scales for each model. A large scaling ra
learned, while a large standard deviation implies inconsistent-scale depth estimation. Compared to th
models combining the MR loss and depth consistency learned much smaller depth values.

configuration of the architecture

$\delta < 1.25^3$	**FPS**
0.982	**127.07**
0.983	62.94
0.982	46.86
0.978	103.44
0.982	68.92
0.980	54.04
0.983	52.61
0.981	51.37
0.984	42.48
0.982	37.79



atio indicates small depth values
ne ground-truth depth values, the