Paper

Unsupervised Monocular Depth Estimation with Left-Right Consistency

,Presented by
Mohamed Nasser

March 9, 2021

Content

- Introduction
- Method
- loss
- Implementation
- Results
- Reproduced Results
- Conclusion

Introduction

- Depth estimation from single image has a long history in computer vision.
- most of the techniques rely on the assumption that multiple observations of the scene of interest are available.
- this work aim to apply novel training objective that enables a convectional neural network to learn to perform single image depth estimation,with absence of ground truth depth

Method

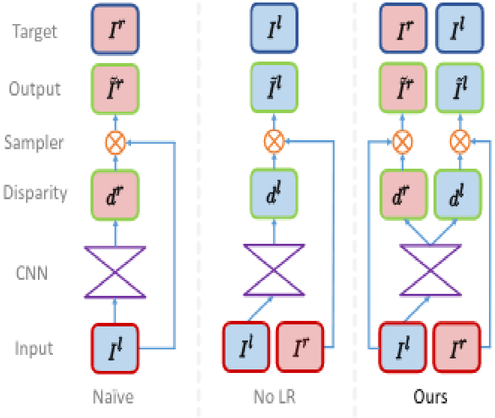
- 1.Depth Estimation as Image Reconstruction
- 2-Depth Estimation Network
- 3-Training Loss

Depth Estimation as Image Reconstruction

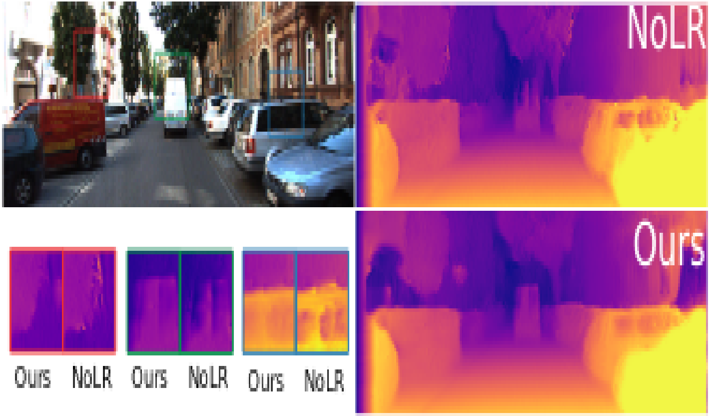
- d=f(I)
- pose depth estimation as an image reconstruction problem during training
- given a calibrated pair of binocular cameras, if we can learn a function that is able to reconstruct one image from the other
- $I^l(d^r) as \mathrm{I\`r}$
- Given the baseline distance between the cameras and the camera focal length

Depth Estimation Network

- Sampling strategies for mapping



Depth Estimation Network



-

Training Loss

- $$C_s = \alpha_{ap}(C^l_{ap} + C^r_{a\,p}) + \alpha_{d\,s}(C^l_{d\,s} + C^r_{d\,s}) + \alpha_{l\,r}(C^l_{l\,r} + C^r_{l\,r})$$

- $$C^l_{ap} = \frac{1}{N}\sum_{i,j}\alpha\frac{1 - SSM(I^l_{i\,j}, I'^l_{ij})}{2} + (1+\alpha)||I^l_{i\,j} - I'^l_{ij}||$$

- $$C^l_{l\,r} = \frac{1}{N}\sum_{i,j}|d^l_{i\,j} - d^r_{i_j + d^l_{i\,j}}|$$

Implementation

- Tensorflow
- Resnet50
- Post-Processing to reduce the effect of stereo dis-occlusions which create disparity ramps on both the left side of the image
- $d_l^{'}$

, flipping back the disparity map : $d_l^{''}$
align to d_l

- disparity maps: the first 5% on the left of the image using $d_l^{''}$ and the last 5% on the right to the disparities from d_l.
The central part of the final disparity map is the average of $d_l and \mathrm{d}_l^{'}$

Results

Method	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	*D1-all*	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours with Deep3D [53]	K	0.412	16.37	13.693	0.512	66.85	0.690	0.833	0.891
Ours with Deep3Ds [53]	K	0.151	1.312	6.344	0.239	59.64	0.781	0.931	0.976
Ours No LR	K	0.123	1.417	6.315	0.220	30.318	0.841	0.937	0.973
Ours	K	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
Ours	CS	0.699	10.060	14.445	0.542	94.757	0.053	0.326	0.862
Ours	CS + K	0.104	1.070	5.417	0.188	25.523	0.875	0.956	0.983
Ours pp	CS + K	0.100	0.934	5.141	0.178	25.077	0.878	0.961	**0.986**
Ours resnet pp	CS + K	**0.097**	**0.896**	**5.093**	**0.176**	**23.811**	**0.879**	**0.962**	**0.986**
Ours Stereo	K	0.068	0.835	4.392	0.146	9.194	0.942	0.978	0.989

Lower is better

Higher is better

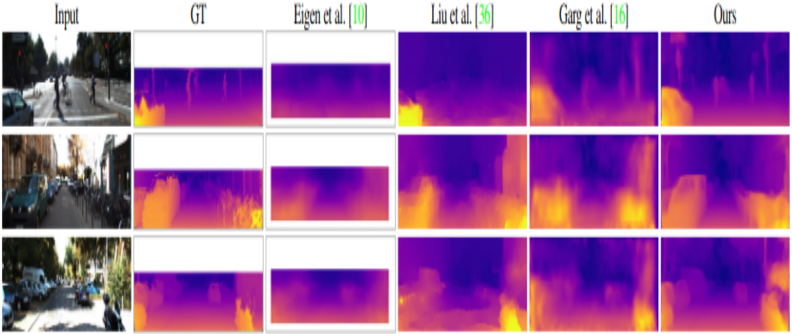
Results

Method	Supervised	Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	No	K	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Eigen et al. [10] Coarse °	Yes	K	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen et al. [10] Fine °	Yes	K	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [36] DCNF-FCSP FT *	Yes	K	0.201	1.584	6.471	0.273	0.68	0.898	0.967
Ours No LR	No	K	0.152	1.528	6.098	0.252	0.801	0.922	0.963
Ours	No	K	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Ours	No	CS + K	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Ours pp	No	CS + K	0.118	0.923	5.015	0.210	0.854	0.947	**0.976**
Ours resnet pp	No	CS + K	**0.114**	**0.898**	**4.935**	**0.206**	**0.861**	**0.949**	**0.976**
Garg et al. [16] L12 Aug 8× cap 50m	No	K	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Ours cap 50m	No	K	0.140	0.976	4.471	0.232	0.818	0.931	0.969
Ours cap 50m	No	CS + K	0.117	0.762	3.972	0.206	0.860	0.948	0.976
Ours pp cap 50m	No	CS + K	0.112	0.680	3.810	0.198	0.866	0.953	**0.979**
Ours resnet pp cap 50m	No	CS + K	**0.108**	**0.657**	**3.729**	**0.194**	**0.873**	**0.954**	**0.979**
Our pp uncropped	No	CS + K	0.134	1.261	5.336	0.230	0.835	0.938	0.971
Ours resnet pp uncropped	No	CS + K	0.130	1.197	5.222	0.226	0.843	0.940	0.971

Lower is better

Higher is better

Results



Results

conclusion

unsupervised deep neural network for single image depth estimation. Instead of using aligned ground truth depth data, which is both rare and costly, binocular stereo data can be captured. this novel loss function enforces consistency between the predicted depth maps from each camera view during training, improving predictions , results are better to fully supervised techniques, which is encouraging for future research that doesn't require expensive to capture ground truth depth