On Regression Losses for Deep Depth Estimation

Marcela Carvalho[1], Bertrand Le Saux[1], Pauline Trouvé-Peloux[1], Andrés Almansa[2], Frédéric Champagnat[1]

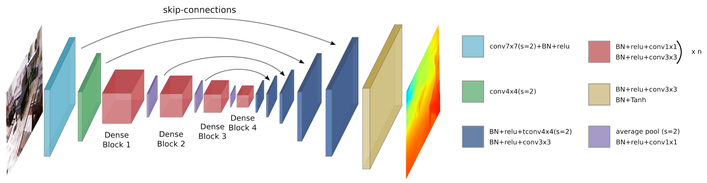
ONERA/DTIS[1], Université Paris Descartes[2]

Motivation

➥ Previous works usually propose both new architecture and regression losses;
➥ Few comparisons are made between loss functions using same architecture.
We propose:
➥ D3-Net: new architecture based on the reuse of previous feature maps that achieves top results on depth estimation (code on github);
➥ End-to-end conditional-GAN for depth estimation based on [1];
➥ Different experiments to compare the performances of the regression losses from the state-of-art.

Network Architecture

D3-Net is an encoder-decoder based on U-Net and DenseNet-121 for the encoder.



➥ The reuse of previous feature maps improve information flow during learning;
➥ Skip and dense connections easy gradient back-propagation to the bottom layers reducing vanishing gradient problems.

Patch-GAN

➥ We adapt the conditional patch GAN proposed in [1] to depth estimation using the LSGAN and a smaller discriminator;
➥ Instead of classifying if the entire image is True/False, the discriminative network classifies by patches.

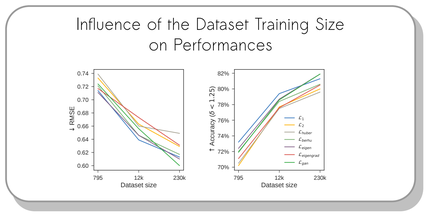
Regression Losses

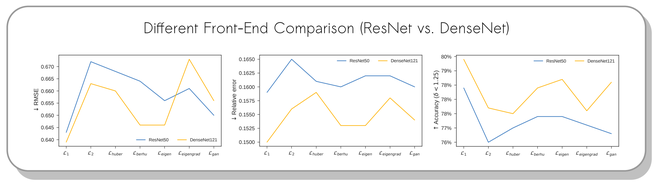
Our experiments make use of the following common losses for regression:

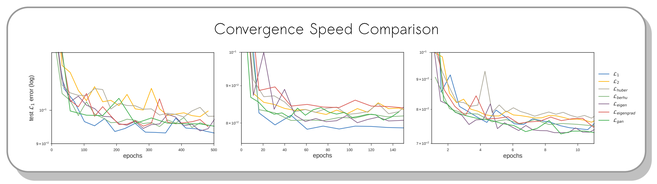
Loss		Equation
Mean absolute	$\mathcal{L}_1$	$\frac{1}{N}\sum_i^N \mid l_i \mid$
Mean square	$\mathcal{L}_2$	$\frac{1}{N}\sum_i^N (l_i)^2$
SI loss [2]	$\mathcal{L}_{eigen}$	$\frac{1}{N}\sum_i^N d_i^2 - \frac{\lambda}{N^2}(\sum_i^N d_i)^2$
SI loss with gradients [3]	$\mathcal{L}_{eigengrad}$	$\frac{1}{N}\sum_i^N d_i^2 - \frac{\lambda}{2N^2}(\sum_i^N d_i)^2 + \frac{1}{N}\sum_i^N[(\nabla_x di)^2 + (\nabla_y di)^2]$
BerHu [4]	$\mathcal{L}_{berhu}$	$\begin{cases} \mathcal{L}_1(l_i) & \mathcal{L}_1(l_i) \leqslant c, \\ \frac{\mathcal{L}_2(l_i)+c^2}{2c} & else. \end{cases}$
Huber [4]	$\mathcal{L}_{huber}$	$\begin{cases} \mathcal{L}_1(l_i) & \mathcal{L}_1(l_i) \geqslant c, \\ \frac{\mathcal{L}_2(l_i)+c^2}{2c} & else. \end{cases}$
LSGAN [5, 1]	$\mathcal{L}_{gan}$	$\frac{1}{2}\mathbb{E}_{x,y\sim p_{data}(x,y)}[(D(x,y) - 1)^2] + \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x,G(x)) - C)^2] + \lambda\mathcal{L}_{L1}(G(x)))$

Let y_i and $\hat{y}_i$ be the ground truth and the estimated distance in meters, $l_i = y_i - \hat{y}_i$, $d_i = log(y_i) - log(\hat{y}_i)$, G, the generator network, D, the discriminator network and x, the input image.

Experiments and Results







Methods	Error↓				Accuracy↑		
	rel	log10	rms	rmslog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Eigen [3]	0.158	-	0.641	0.214	76.9%	95.0%	98.8%
Laina [4]	0.127	0.055	0.573	0.195	81.1%	95.3%	98.8%
Xu [6]	0.121	0.052	0.586	-	81.1%	95.4%	98.7%
Jung[7]	0.134	-	0.527	-	82.2%	97.1%	99.3%
Kendall and Gal [8]	0.110	0.045	0.506	-	81.7%	95.9%	98.9%
D3-Net*	0.136	-	0.504	-	82.1%	95.5%	98.7%

*Results were updated from original paper.

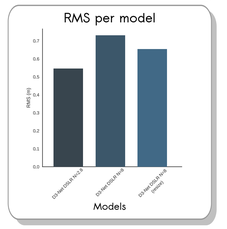
➥ ↑ data leads to better results;
➥ Losses evolve differently when increasing available data;
➥ $\mathcal{L}_{gan}$ becomes highly efficient with more data, which helps on the stability of the method.
➥ DenseNet gets better performances compared to ResNet;
➥ $\mathcal{L}_1$ and $\mathcal{L}_{eigen}$ show better results for the ensemble of the experiments;
➥ Best results using $\mathcal{L}_{gan}$ show better segmented images and more object details.

Work

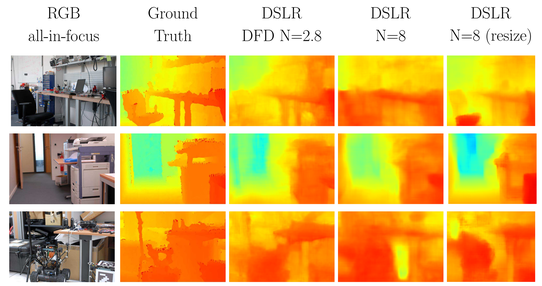
In [9], we explore D3-Net to perform depth estimation with the presence of defocus blur.

Deep Depth-from-Defocus (DFD) with real data

DSLR + Asus Xtion

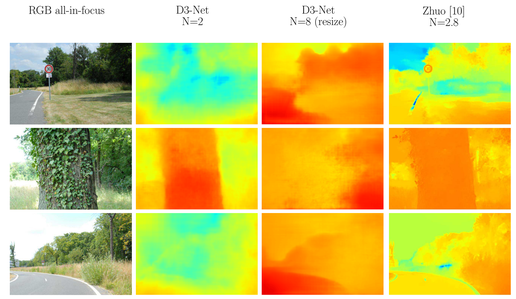


We create a new dataset with a platform which contains a DSLR camera and a RGB-D sensor.
➥ Pretrain on defocused NYUv2;
➥ Finetune on real data.



Deep-DFD in the wild with outdoor scenes

We now observe the network generalization to outdoor challenging scenes.



Conclusions:
➥ Combined information of geometrical structure and defocus blur avoids classical limitations of DFD techniques;
➥ Deep-DFD is a promissing method to generalize learning depth estimation;
➥ Co-conception of a sensor and a deep depth estimation methods.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *NIPS*, 2014.

[3] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture," *ICCV*, 2015.

[4] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*, IEEE, 2016, pp. 239–248.

[5] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," *arXiv preprint ArXiv:1611.04076*, 2016.

[6] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," *arXiv preprint arXiv:1704.02157*, 2017.

[7] H. Jung, Y. Kim1, D. Min, C. Oh, and K. Sohn, "Depth prediction from a single image with conditional adversarial networks," in *ICIP*, 2017.

[8] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[9] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: how can defocus blur improve 3D estimation using dense neural networks?" *3DRW ECCV Workshop*, 2018.

[10] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recognition*, 2011.

github.com/marcelampc/d3net_depth_estimation