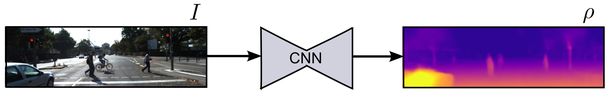
Semi-Supervised Deep Learning for Monocular Depth Map Prediction

Yevhen Kuznietsov, Jörg Stückler, Bastian Leibe
Computer Vision Group, Visual Computing Institute, RWTH Aachen University

IEEE 2017 Conference on Computer Vision and Pattern Recognition

Problem Statement

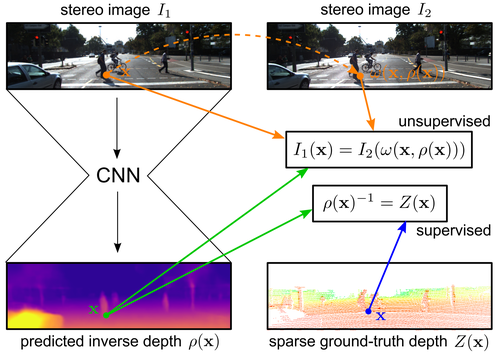
Task: Infer dense depth from monocular RGB images



Challenge:
- Supervised learning methods require ground truth depth
- Dense high-quality ground-truth depth is difficult to obtain
- Sensors such as 3D lasers measure sparsely and have own error characteristics

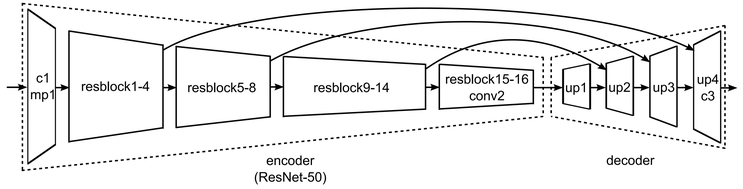
Our Approach

Approach: complement supervised learning with unsupervised cues for depth prediction

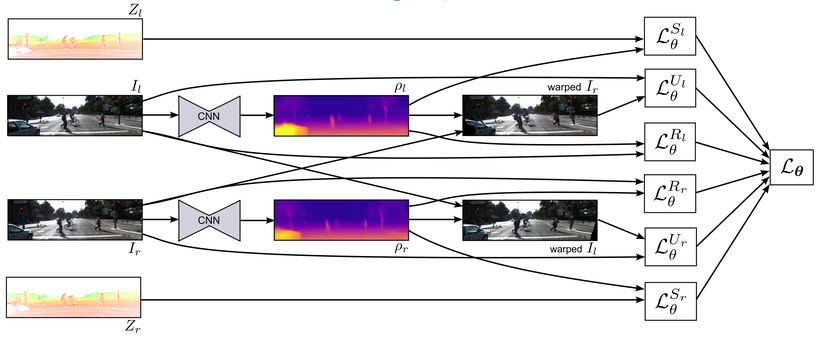


Warping: $\omega(\mathbf{x}, \rho(\mathbf{x})) := \mathbf{x} - f\, b\, \rho(\mathbf{x})$

Network Architecture



Training Pipeline



Both frames in stereo pair are treated equally. This enables accurrate depth predictions at image borders without postprocessing.

$$\mathcal{L}_{\boldsymbol{\theta}}\left(I_l, I_r, Z_l, Z_r\right) = \lambda_t \mathcal{L}_{\boldsymbol{\theta}}^S\left(I_l, I_r, Z_l, Z_r\right) + \gamma \mathcal{L}_{\boldsymbol{\theta}}^U\left(I_l, I_r\right) + \mathcal{L}_{\boldsymbol{\theta}}^R\left(I_l, I_r\right)$$

Supervised:

$$\mathcal{L}_{\boldsymbol{\theta}}^S = \sum_{\mathbf{x}\in\Omega_{Z,l}} \left\|\rho_{l,\boldsymbol{\theta}}(\mathbf{x})^{-1} - Z_l(\mathbf{x})\right\|_\delta + \sum_{\mathbf{x}\in\Omega_{Z,r}} \left\|\rho_{r,\boldsymbol{\theta}}(\mathbf{x})^{-1} - Z_r(\mathbf{x})\right\|_\delta$$

$$\|d\|_\delta = \begin{cases} |d|, d \le \delta \\ \frac{d^2+\delta^2}{2\delta}, d > \delta \end{cases} \qquad \delta = 0.2\, \max_{\mathbf{x}\in\Omega_Z}\left(\left|\rho(\mathbf{x})^{-1} - Z(\mathbf{x})\right|\right)$$

Unsupervised:

$$\mathcal{L}_{\boldsymbol{\theta}}^U = \sum_{\mathbf{x}\in\Omega_{U,l}} \left|(\mathbf{G}_\sigma * I_l)(\mathbf{x}) - (\mathbf{G}_\sigma * I_r)(\omega(\mathbf{x}, \rho_{l,\boldsymbol{\theta}}(\mathbf{x})))\right| + \sum_{\mathbf{x}\in\Omega_{U,r}} \left|(\mathbf{G}_\sigma * I_r)(\mathbf{x}) - (\mathbf{G}_\sigma * I_l)(\omega(\mathbf{x}, -\rho_{r,\boldsymbol{\theta}}(\mathbf{x})))\right|$$

Regularization:

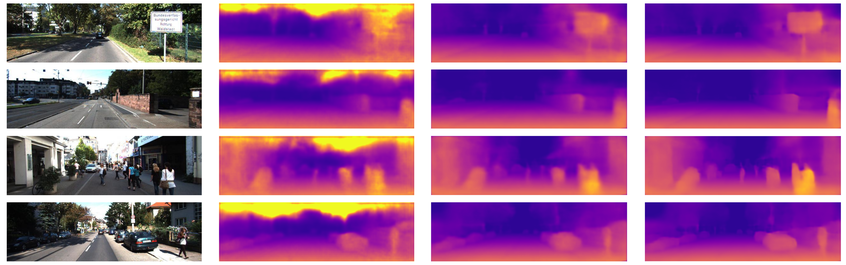
$$\mathcal{L}_{\boldsymbol{\theta}}^R = \sum_{i\in\{l,r\}} \sum_{\mathbf{x}\in\Omega} \left|\phi\left(\nabla I_i(\mathbf{x})\right)^\top \nabla \rho_i(\mathbf{x})\right|$$

$$\phi(\mathbf{g}) = \left(\exp(-\eta\,|g_x|), \exp(-\eta\,|g_y|)\right)^\top \text{ and } \eta = \tfrac{1}{255}$$

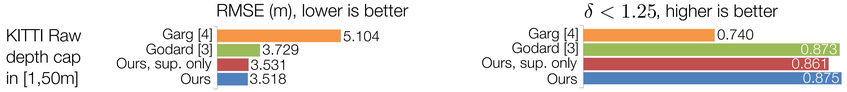
Setup

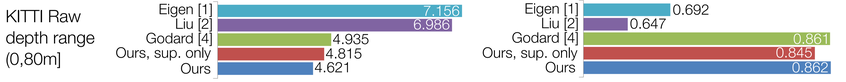
Dataset: KITTI Raw. Test set: 697 images (28 sequences) used by Eigen et al. [1]
Training/validation sets sampled from remaining 33 sequences, resulting in 7346/244 images
Optimization: SGD with momentum. Weight decay. Dropout p = 0.5 at last up-projection layer
Data augmentation: Gamma and brightness in range [0.8; 1.2]
I/O resolution: 621 x 187 px >> 320 x 96 px
Initialization: Encoder pretrained on ImageNet, variance scaling initialization of decoder
Hardware: 6 GB NVIDIA GTX 980Ti (batch size 5)

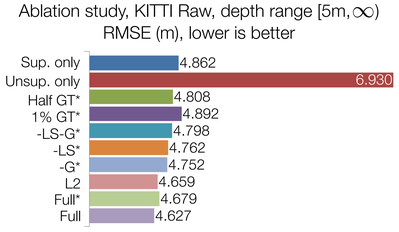
Results

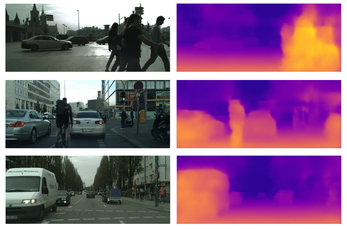


KITTI Raw. RGB | Supervised only | Ours, half GT | **Ours, full**









Cityscapes results, model trained on KITTI.

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] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. CVPR, 2015.
[3] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. CVPR, 2017.
[4] R. Garg, V. Kumar, G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. ECCV, 2016

This work has been supported by European Research Council Starting Grant CV-SUPER (ERC-2012-StG-307432).