Semi-Supervised Deep Learning for Monocular Depth Map Prediction

Yevhen Kuznietsov          Jörg Stückler          Bastian Leibe
Computer Vision Group, Visual Computing Institute, RWTH Aachen University
52074 Aachen, Germany
yevhen.kuznietsov@rwth-aachen.de, stueckler | leibe @vision.rwth-aachen.de

Abstract

Supervised deep learning often suffers from the lack of sufficient training data. Specifically in the context of monocular depth map prediction, it is barely possible to determine dense ground truth depth images in realistic dynamic outdoor environments. When using LiDAR sensors, for instance, noise is present in the distance measurements, the calibration between sensors cannot be perfect, and the measurements are typically much sparser than the camera images. In this paper, we propose a novel approach to depth map prediction from monocular images that learns in a semi-supervised way. While we use sparse ground-truth depth for supervised learning, we also enforce our deep network to produce photoconsistent dense depth maps in a stereo setup using a direct image alignment loss. In experiments we demonstrate superior performance in depth map prediction from single images compared to the state-of-the-art methods.

1. Introduction

Estimating depth from single images is an ill-posed problem which cannot be solved directly from bottom-up geometric cues in general. Instead, a-priori knowledge about the typical appearance, layout and size of objects needs to be used, or further cues such as shape from shading or focus have to be employed which are difficult to model in realistic settings. In recent years, supervised deep learning approaches have demonstrated promising results for single image depth prediction. These learning approaches appear to capture the statistical relationship between appearance and distance to objects well.

Supervised deep learning, however, requires vast amounts of training data in order to achieve high accuracy and to generalize well to novel scenes. Supplementary depth sensors are typically used to capture ground truth. In the indoor setting, active RGB-D cameras can be used. Outdoors, 3D laser scanners are a popular choice to capture depth measurements. However, using such sensing devices bears several shortcomings. Firstly, the sensors have their own error and noise characteristics, which will be learned by the network. In addition, when using 3D lasers, the measurements are typically much sparser than the images and do not capture high detail depth variations visible in the images well. Finally, accurate extrinsic and intrinsic calibration of the sensors is required. Ground truth data could alternatively be generated through synthetic rendering of depth maps. The rendered images, however, do not fully realistically display the scene and do not incorporate real image noise characteristics.

Very recently, unsupervised methods have been introduced [6, 9] that learn to predict depth maps directly from the intensity images in a stereo setup—without the need for an additional supplementary modality for capturing the ground truth. One drawback of these approaches is the well-
known fact that stereo depth reconstruction based on image matching is an ill-posed problem on its own. To this end, common regularization schemes can be used which impose priors on the depth such as small depth gradient norms which may not be fully satisfied in the real environment.

In this paper, we propose a semi-supervised learning approach that makes use of supervised as well as unsupervised training cues to incorporate the best of both worlds. Our method benefits from ground-truth measurements as an unambiguous (but noisy and sparse) cue for the actual depth in the scene. Unsupervised image alignment complements the ground-truth by a huge amount of additional training data which is much simpler to obtain and counteracts the deficiencies of the ground-truth depth measurements. By the combination of both methods, we achieve significant improvements over the state-of-the-art in single image depth map prediction which we evaluate on the popular KITTI dataset [7] in urban street scenes. We base our approach on a state-of-the-art deep residual network in an encoder-decoder architecture for this task [16] and augment it with long skip connections between corresponding layers in encoder and decoder to predict high detail output depth maps. Our network converges quickly to a good model from little supervised training data, mainly due to the use of pretrained encoder weights (on ImageNet [22] classification task) and unsupervised training. The use of supervised training also simplifies unsupervised learning significantly. For instance, a tedious coarse-to-fine image alignment loss as in previous unsupervised learning approaches [6] is not required in our semi-supervised approach.

In summary, we make the following contributions: 1) We propose a novel semi-supervised deep learning approach to single image depth map prediction that uses supervised as well as unsupervised learning cues. 2) Our deep learning approach demonstrates state-of-the-art performance in challenging outdoor scenes on the KITTI benchmark.

2. Related Work

Over the last years, several learning-based approaches to single image depth reconstruction have been proposed that are trained in a supervised way. Often, measured depth from RGB-D cameras or 3D laser scanners is used as ground-truth for training. Saxena et al. [24] proposed one of the first supervised learning-based approaches to single image depth map prediction. They model depth prediction in a Markov random field and use multi-scale texture features that have been hand-crafted. The method also combines monocular cues with stereo correspondences within the MRF.

Many recent approaches learn image features using deep learning techniques. Eigen et al. [5] propose a CNN architecture that integrates coarse-scale depth prediction with fine-scale prediction. The approach of Li et al. [17] combines deep learning features on image patches with hierarchical CRFs defined on a superpixel segmentation of the image. They use pretrained AlexNet [14] features of image patches to predict depth at the center of the superpixels. A hierarchical CRF refines the depth across individual pixels. Liu et al. [20] also propose a deep structured learning approach that avoids hand-crafted features. Their deep convolutional neural fields allow for training CNN features of unary and pairwise potentials end-to-end, exploiting continuous depth and Gaussian assumptions on the pairwise potentials. Very recently, Laina et al. [16] proposed to use a ResNet-based encoder-decoder architecture to produce dense depth maps. They demonstrate the approach to predict depth maps in indoor scenes using RGB-D images for training. Further lines of research in supervised training of depth map prediction use the idea of depth transfer from example images [13, 12, 21], or integrate depth map prediction with semantic segmentation [15, 19, 4, 26, 18].

Only few very recent methods attempt to learn depth map prediction in an unsupervised way. Garg et al. [6] propose an encoder-decoder architecture similar to FlowNet [3] which is trained to predict single image depth maps on an image alignment loss. The method only requires images of a corresponding camera in a stereo setup. The loss quantifies the photometric error of the input image warped into its corresponding stereo image using the predicted depth. The loss is linearized using first-order Taylor approximation and hence requires coarse-to-fine training. Xie et al. [27] do not regress the depth maps directly, but produce probability maps for different disparity levels. A selection layer then reconstructs the right image using the left image and these probability maps. The network is trained to minimize pixel-wise reconstruction error. Godard et al. [9] also use an image alignment loss in a convolutional encoder-decoder architecture but additionally enforce left-right consistency of the predicted disparities in the stereo pair. Our semi-supervised approach simplifies the use of unsupervised cues and does not require multi-scale depth map prediction in our network architecture. We also do not explicitly enforce left-right consistency, but use both images in the stereo pair equivalently to define our loss function. The semi-supervised method of Chen et al. [1] incorporates the side-task of depth ranking of pairs of pixels for training a CNN on single image depth prediction. For the ranking task, ground-truth is much easier to obtain but only indirectly provides information on continuous depth values. Our approach uses image alignment as a geometric cue which does not require manual annotations.

3. Approach

We base our approach on supervised as well as unsupervised principles for learning single image depth map prediction (see Fig. [1]). A straight-forward approach is to use a supplementary measuring device such as a 3D laser
in order to capture ground-truth depth readings for supervised training. This process typically requires an accurate extrinsic calibration between the 3D laser sensor and the camera. Furthermore, the laser measurements have several shortcomings. Firstly, they are affected by erroneous readings and noise. They are also typically much sparser than the camera images when projected into the image. Finally, the center of projection of laser and camera do not coincide. This causes depth readings of objects that are occluded from the viewpoint of the camera to project into the camera image. To counteract these drawbacks, we make use of two-view geometry principles to learn depth prediction directly from the stereo camera images in an unsupervised way. We achieve this by direct image alignment of one stereo image to the other. This process only requires a known camera calibration and the depth map predicted by the CNN. Our semi-supervised approach learns from supervised and unsupervised cues concurrently.

We train the CNN to predict the inverse depth $\rho(x)$ at each pixel $x \in \Omega$ from the RGB image $I$. According to the ground truth, the predicted inverse depth should correspond to the LiDAR depth measurement $Z(x)$ that projects to the same pixel, i.e.,

$$\rho(x)^{-1} = Z(x).$$  \hspace{1cm} (1)

However, the laser measurements only project to a sparse subset $\Omega_Z \subseteq \Omega$ of the pixels in the image.

As the unsupervised training signal, we assume photo-consistency between the left and right stereo images, i.e.,

$$I_1(x) \overset{\lambda}{=} I_2(\omega(x, \rho(x))).$$  \hspace{1cm} (2)

In our calibrated stereo setup, the warping function can be defined as

$$\omega(x, \rho(x)) := x - fb\rho(x)$$  \hspace{1cm} (3)

on the rectified images, where $f$ is the focal length and $b$ is the baseline. This image alignment constraint holds at every pixel in the image.

We additionally make use of the interchangeability of the stereo images. We quantify the supervised loss in both images by projecting the ground truth laser data into each of the stereo images. We also constrain the depth estimate between the left and right stereo images to be consistent implicitly by enforcing photometric consistency based on the inverse depth prediction for both images, i.e.,

$$I_{left}(x) \overset{1}{=} I_{right}(\omega(x, \rho(x)))$$

$$I_{right}(x) \overset{1}{=} I_{left}(\omega(x, -\rho(x))).$$  \hspace{1cm} (4)

Finally, in textureless regions without ground truth depth readings, the depth map prediction problem is ill-posed and an adequate regularization needs to be imposed.

3.1. Loss function

We formulate a single loss function that incorporates both types of constraints that arise from supervised and unsupervised cues seamlessly,

$$L_\theta(I_1, I_2, Z_1, Z_2) =$$

$$\lambda I L_S^\theta(I_1, I_2, Z_1, Z_2) + \gamma L_U^\theta(I_1, I_2) + L_R^\theta(I_1, I_2),$$  \hspace{1cm} (5)

where $\lambda$ and $\gamma$ are trade-off parameters between supervised loss $L_S^\theta$, unsupervised loss $L_U^\theta$, and a regularization term $L_R^\theta$.

Supervised loss. The supervised loss term measures the deviation of the predicted depth map from the available
ground truth at the pixels,
\[ \mathcal{L}_d^S = \sum_{x \in \Omega_{1,1}} \| \rho_{l, \theta}(x)^{-1} - Z_l(x) \|_\delta 
\]
\[ + \sum_{x \in \Omega_{1,1}} \| \rho_{r, \theta}(x)^{-1} - Z_r(x) \|_\delta. \]  

We use the berHu norm \( \| \cdot \|_\delta \) as introduced in [16] to focus training on larger depth residuals during CNN training,
\[ \| d \|_\delta = \begin{cases} |d|, & d \leq \delta \\
\frac{d^2 + \delta^2}{2\delta}, & d > \delta \end{cases}. \]  

We adaptively set
\[ \delta = 0.2 \max_{x \in \Omega_{1,1}} \left( |\rho(x)^{-1} - Z(x)| \right). \]  

**Unsupervised loss.** The unsupervised part of our loss quantifies the direct image alignment error in both directions
\[ \mathcal{L}_d^U = \sum_{x \in \Omega_{1,1}} \| (G_{\sigma} \ast I_l)(x) - (G_{\sigma} \ast I_r)(\omega(x, \rho_{l, \theta}(x))) \|_2 + \sum_{x \in \Omega_{1,1}} \| (G_{\sigma} \ast I_r)(x) - (G_{\sigma} \ast I_l)(\omega(x, -\rho_{r, \theta}(x))) \|_2, \]  

with a Gaussian smoothing kernel \( G_{\sigma} \) with a standard deviation of \( \sigma = 1 \) px. We found this small amount of Gaussian smoothing to be beneficial, presumably due to reducing image noise. We evaluate the direct image alignment loss at the sets of image pixels \( \Omega_{1,1} \) of the reconstructed images that warp to a valid location in the second image.

**Regularization loss.** As suggested in [9], the smoothness term penalizes depth changes at pixels with low intensity variation. In order to allow for depth discontinuities at object contours, we downscale the regularization term anisotropically according to the intensity variation:
\[ L_d^R = \sum_{i \in \{l, r\}} \sum_{x \in \Omega_{1,1}} \phi(\nabla I_i(x)) \times \nabla \rho_{i, \theta}(x) \]  

with \( \phi(\mathbf{g}) = (\exp(-\eta |g_x|), \exp(-\eta |g_y|))^T \) and \( \eta = \frac{1}{255} \).

Supervised, unsupervised, and regularization terms are seamlessly combined within our novel semi-supervised loss function formulation (see Fig. 2). In contrast to previous methods, our approach treats both cameras in a stereo setup equivalently. All three loss components are formulated in a symmetric way for the cameras which implicitly enforces consistency in the predicted depth maps between the cameras.

| Layer | Channels I/O | Scaling | Inputs |
|-------|--------------|---------|--------|
| conv1 \_ | 3 / 64 | 2 | RGB |
| max_pool1 \_ | 64 / 64 | 4 | conv1 |
| res_block1 \_ | 64 / 256 | 4 | max_pool1 |
| res_block2 \_ | 256 / 256 | 4 | res_block1 |
| res_block3 \_ | 256 / 256 | 4 | res_block2 |
| res_block4 \_ | 512 / 512 | 8 | res_block3 |
| res_block5 \_ | 512 / 512 | 8 | res_block4 |
| res_block6 \_ | 512 / 512 | 8 | res_block5 |
| res_block7 \_ | 512 / 512 | 8 | res_block6 |
| res_block8 \_ | 512 / 1024 | 16 | res_block7 |
| res_block9 \_ | 1024 / 1024 | 16 | res_block8 |
| res_block10 \_ | 1024 / 1024 | 16 | res_block9 |
| res_block11 \_ | 1024 / 1024 | 16 | res_block10 |
| res_block12 \_ | 1024 / 1024 | 16 | res_block11 |
| res_block13 \_ | 1024 / 1024 | 16 | res_block12 |
| res_block14 \_ | 1024 / 2048 | 32 | res_block13 |
| res_block15 \_ | 2048 / 2048 | 32 | res_block14 |
| res_block16 \_ | 2048 / 2048 | 32 | res_block15 |
| conv2 \_ | 2048 / 2048 | 32 | res_block16 |
| upproject1 | 1024 / 512 | 16 | conv2 |
| upproject2 | 512 / 256 | 8 | upproject1 |
| upproject3 | 256 / 128 | 4 | upproject2 |
| upproject4 | 128 / 64 | 2 | upproject3 |
| conv3 \_ | 64 / 1 | 2 | upproject4 |

Table 1. Layers in our deep residual encoder-decoder architecture. We input the final output layers at each resolution of the encoder at the respective decoder layers (long skip connections). This facilitates the prediction of fine detailed depth maps by the CNN.

### 3.2. Network Architecture

We use a deep residual network architecture in an encoder-decoder scheme, similar to the supervised approach in [16] (see Fig. 3). Taking inspiration from non-residual architectures such as FlowNet [3], our architecture includes long skip connections between the encoder and decoder to facilitate fine detail predictions at the output resolution. Table 1 details the various layers in our network.

Input to our network is the RGB camera image. The encoder resembles a ResNet-50 [11] architecture (without the final fully connected layer) and successively extracts low-
residual high-dimensional features from the input image. The encoder subsamples the input image in 5 stages, the first stage convolving the image to half input resolution and each successive stage stacking multiple residual blocks. The decoder upprojects the output of the encoder using residual blocks. We found that adding long skip-connections between corresponding layers in encoder and decoder to this architecture slightly improves the performance on all metrics without affecting convergence. Moreover, the network is able to predict more detailed depth maps than without skip connections.

We denote a convolution of filter size $k \times k$ and stride $s$ by $\text{conv}^s_k$. The same notation applies to pooling layers, e.g., $\text{max\_pool}^s_k$. Each convolution layer is followed by batch normalization with exception of the last layer in the network. Furthermore, we use ReLU activation functions on the output of the convolutions except at the inputs to the sum operation of the residual blocks where the ReLU comes after the sum operation. resblock$_i^s$ denotes the residual block of type $i$ with stride $s$ at its first convolution layer, see Figs. 4 and 5 for details on each type of residual block. Smaller feature blocks consist of $16s$ maps, while larger blocks contain 4 times more feature maps, where $s$ is the output scale of the residual block. Lastly, upproject is the upprojection layer proposed by Laina et al. [16]. We use the fast implementation of upprojection layers, but for better illustration we visualize upprojection by its "naive" version (see Fig. 6).

4. Experiments

We evaluate our approach on the raw sequences of the KITTI benchmark [7] which is a popular dataset for single image depth map prediction. The sequences contain stereo imagery taken from a driving car in an urban scenario. The dataset also provides 3D laser measurements from a Velodyne laser scanner that we use as ground-truth measurements. This dataset has been used to train and evaluate the state-of-the-art methods and allows for quantitative comparison.

We use the split into 28 training and 28 testing scenes as proposed by Eigen et al. [5]. For training we decided to even the sequence distribution with 450 frames per sequence. This results in 7346 unique frames and 12600 frames in total for training. Since there is no official validation set, we create one by sampling every tenth frame from the remaining 5 sequences with little image motion. All these sequences are urban, so we additionally select those frames from the training sequences that are in the middle between 2 training images with distance of at least 20 frames. In total we obtain a validation set of 100 urban and
4.1. Implementation Details

We initialize the encoder part of our network with ResNet-50 [11] weights pretrained for ImageNet classification task. The convolution filter weights in the decoder part are initialized randomly according to the approach of Glorot and Bengio [8]. We also tried the initialization by He et al. [10] but did not notice any performance difference. We predict the inverse depth and initialize the network in such a way that the predicted values are close to 0 in the beginning of training. This way, the unsupervised direct image alignment loss is initialized with almost zero disparity between the images. However, this also results in large gradients from the supervised loss which would cause divergence of the model. To achieve a convergent optimization, we slowly fade-in the supervised loss with the number of iterations using $\lambda_t = \beta e^{-\frac{t}{T}}$. We also experimented with gradually fading in the unsupervised loss, but experienced degraded performance on the upper part of the image. In order to avoid overfitting we use L2 regularization on all the model weights with weight decay $w_d = 0.00004$. We also apply dropout to the output of the last up-projection layer with a dropout probability of 0.5.

To train the CNN on KITTI we use stochastic gradient descent with momentum with a learning rate of 0.01 and momentum of 0.9. We train the variants of our model for at least 15 epochs on a 6 GB NVIDIA GTX 980Ti with 6 GB memory which allows for a batch size of 5. We stop training when the validation loss starts to increase and select the model with the best RMSE score on the validation set. The network is trained on a resolution of 621 × 187 pixels for both input images and ground truth depth maps. Hence, the resolution of the predicted inverse depth maps is 320 × 96. For evaluation we upsample the predicted depth maps to the resolution of the ground truth. For data augmentation, we use $\gamma$-a augmentation and also randomly multiply the intensities of the input images by a value $\alpha \in [0.8; 1.2]$. We use linear interpolation for subpixel-level warping. The inference from one image takes 0.048 s in average.

4.2. Evaluation Metrics

We evaluate the accuracy of our method in depth prediction using the 3D laser ground truth on the test images. We use the following depth evaluation metrics used by Eigen et al. [5]:

RMSE: $\sqrt{\frac{1}{T} \sum_{i=1}^{T} \left\| \rho(x_i)^{-1} - Z(x_i) \right\|^2}$.

RMSE (log): $\sqrt{\frac{1}{T} \sum_{i=1}^{T} \left\| \log(\rho(x_i)^{-1}) - \log(Z(x_i)) \right\|^2}$,

\[
\left\{ i \in \{1, \ldots, T\} \mid \max_{\delta < \text{thr}} \frac{Z(x_i)}{\rho(x_i)^{-1}} = \delta < \text{thr} \right\}.
\]

Accuracy: $\frac{1}{T} \sum_{i=1}^{T} \frac{1}{\rho(x_i)^{-1}} - Z(x_i) \right\|^2}$.

where $T$ is the number of pixels with ground-truth in the test set.

In order to compare our results with Eigen et al. [5] and Godard et al. [9], we crop our image to the evaluation crop applied by Eigenet al. (crop A). We also use the same resolution of the ground truth depth image and cap the predicted depths at 80 m [9]. For comparison with Garg et al. [6], we also provide results when capping the predicted depths at 50 m maximum depth and using their 608 × 176 crop (crop B). For completeness, we also give results for our method evaluated on the uncropped image without a cap on the maximum predicted depth. For all cases the minimum ground-truth depth is 5 m.

4.3. Results

Table 2 shows our results in relation to the state-of-the-art methods on the test images of the KITTI benchmark. For most metrics and setups (all except of $\delta < 1.25$ with 50 m cap), our system clearly performs the best. We outperform the best setup of Godard et al. [9] (unpublished) by more than 17% in terms of RMSE and almost by 22% for its log scale on crop A and cap of 80 m. When evaluating at a prediction cap of 50 m and crop B, our predictions are still in average 19 cm more accurate. However, since there are ground truth points with depth bigger than 50 m present in the dataset and our method is able to provide high-quality predictions for the corresponding pixels, capping the predicted depth makes our approach perform worse. In particular, RMSE is decreased by more than 12 cm for uncapped evaluation at the same crop.

We also qualitatively compare the output of our method with the state-of-the-art in Fig. 7. In some parts, the predictions of Godard et al. [9] may appear more detailed and our depth maps seem to be smoother. However, these details are not always consistent with the ground truth depth maps as also indicated by the quantitative results. For instance, our predictions for the thin traffic poles and lights of the top frame in Figure 7 are more accurate.

We also analyze the contributions of the various design choices in our approach (see Table 3). The computation of the unsupervised loss term on all valid pixels improves the performance comparing to the variant with unsupervised term evaluated only for valid pixels without available ground truth. When using the L2-norm on the supervised loss instead of the berHu norm, the RMSE evaluation metric on the ground-truth depth improves on the validation set, but is worse on the test set. The L2-norm also visually produces noisier depth maps. Thus, we prefer to use BerHu over L2, which reduces the noise (see Fig. 8) and performs
Table 2. Quantitative results of our method and approaches reported in the literature on the test set of the KITTI raw dataset used by Eigen et al. \cite{5} for different depth prediction caps. Best results shown in bold.

| Approach                        | Crop | Cap | RMSE | RMSE (log) | ARD | SRD | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
|---------------------------------|------|-----|------|------------|-----|-----|-----------------|-------------------|-------------------|
| Eigen et al. \cite{5} coarse 28×144 | A    | 80 m| 7.216| 0.273      | 0.228| -   | 0.679          | 0.897             | 0.967             |
| Eigen et al. \cite{5} fine 27×142 | A    | 80 m| 7.156| 0.270      | 0.215| -   | 0.692          | 0.899             | 0.967             |
| Liu et al. \cite{20} DCNF-FCSP FT | A    | 80 m| 6.986| 0.289      | 0.217| 1.841| 0.647          | 0.882             | 0.961             |
| Godard et al. \cite{9} (unpublished) | A    | 80 m| 5.849| 0.242      | 0.141| 1.369| 0.818          | 0.929             | 0.966             |
| Godard et al. \cite{9} + CS (unpublished) | A    | 80 m| 5.763| 0.236      | 0.136| 1.512| 0.836          | 0.935             | 0.968             |
| Ours, supervised only           | A    | 80 m| 5.157| 0.202      | 0.129| 0.679| 0.897          | 0.967             | 0.961             |
| Ours, unsupervised only         | A    | 80 m| 7.323| 0.288      | 0.202| 3.242| 0.766          | 0.919             | 0.964             |
| Ours                           | A    | 80 m| 4.903| 0.194      | 0.123| 0.864| 0.830          | 0.954             | 0.986             |

Table 3. Quantitative results of different variants of our approach on the KITTI raw dataset used by Eigen et al. \cite{5} (without cropping and capping the predicted depth). Approaches marked with * evaluate the unsupervised loss only for the pixels without available ground truth. Best results shown in bold.

| Approach                                      | RMSE | RMSE (log) | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
|-----------------------------------------------|------|------------|-----------------|-------------------|-------------------|
| Our full approach                              |      |            |                 |                   |                   |
| Our full approach *                           | 4.627| 0.189      | 0.856           | 0.960             | 0.986             |
| L2-norm instead of BerHu-norm in supervised loss | 4.679| 0.192      | 0.854           | 0.959             | 0.985             |
| No long skip connections *                     | 4.659| 0.195      | 0.841           | 0.958             | 0.986             |
| No Gaussian smoothing in unsupervised loss *   | 4.762| 0.194      | 0.853           | 0.958             | 0.985             |
| No long skip connections and no Gaussian smoothing * | 4.752| 0.193      | 0.854           | 0.958             | 0.986             |
| Supervised training only                      | 4.798| 0.195      | 0.853           | 0.957             | 0.984             |
| Unsupervised training only (50 m cap)         | 4.862| 0.197      | 0.839           | 0.956             | 0.986             |
| Only 50 % of laser points used *              | 6.930| 0.330      | 0.745           | 0.903             | 0.952             |
| Only 1 % of laser points used *               | 4.808| 0.192      | 0.852           | 0.958             | 0.986             |

better on the test set. We also found that our system benefits from both long skip connections and Gaussian smoothing in the unsupervised loss. The latter also results in slightly faster convergence.

To show that our approach benefits from the semi-supervised pipeline, we also give results for purely supervised and purely unsupervised training. For purely supervised learning, our network achieves less accurate depth map prediction than in the semi-supervised setting. In the unsupervised case, the depth maps include larger amounts of outliers such that we provide results for capped depth predictions at a maximum of 50 m. Here, our network seems to perform less well than the unsupervised methods of Godard et al. \cite{9} and Garg et al. \cite{6}. Notably, our approach does not perform multi-scale image alignment, but uses the available ground truth to avoid local optima of the direct image alignment. We also demonstrate that our system does not suffer severely if the ground truth depth is reduced to 50% or 1% of the available measurements. To this end, we subsample the available laser data prior to projecting it into the camera image.

Our results clearly demonstrate the benefit of using a deep residual encoder-decoder architecture with long skip connection for the task of single image depth map prediction. Our semi-supervised approach gives additional training cues to the supervised loss through direct image alignment. This combination is even capable to improve depth prediction error for the laser ground-truth compared to purely supervised learning. Our semi-supervised learning method converges much faster (in about one third the number of iterations) than purely supervised training.

We also demonstrate the generalization ability of our
Figure 7. Qualitative results and comparison with state-of-the-art methods. Ground-truth (GT) has been interpolated for visualization. Note the crisper prediction of our method on objects such as cars, pedestrians and traffic signs. Also notice, how our method can learn appropriate depth predictions in the upper part of the image that is not covered by the ground-truth.

Figure 8. Qualitative results of variants of our semi-supervised learning approach. Shown variants are the full model (full), without long skip connections (no LS), with L2 norm on the supervised loss (L2) and using half the ground-truth laser measurements (half GT).

Figure 9. Qualitative results on Make3D (left 2) and Cityscapes (right).

5. Conclusions

In this paper, we proposed a novel semi-supervised deep learning approach to monocular depth map prediction. Purely supervised learning requires a vast amount of data. In outdoor environments, often supplementary sensors such as 3D lasers have to be used to acquire training data. These sensors come with their own shortcoming such as specific error and noise characteristics and sparsity of the measurements. We complement such supervised cues with unsupervised learning based on direct image alignment between the images in a stereo camera setup. We quantify the photoconsistency of pixels in both images that correspond to each other according to the depth predicted by the CNN.

We use a state-of-the-art deep residual network in an encoder-decoder architecture and enhance it with long skip connections. Our main contribution is a seamless combination of supervised, unsupervised, and regularization terms in our semi-supervised loss function. The loss terms are defined symmetrically for the available cameras in the stereo setup, which implicitly promotes consistency in the depth estimates. Our approach achieves state-of-the-art performance in single image depth map prediction on the popular KITTI dataset. It is able to predict detailed depth maps on thin and distant objects. It also estimates reasonable depth in image parts in which there is no ground-truth available for supervised learning.

In future work, we will investigate semi-supervised learning for further tasks such as semantic image segmentation. Our approach could also be extended to couple monocular and stereo depth cues in a unified deep learning framework.

References

[1] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In NIPS, 2016.
[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.
[3] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In ICCV, 2015.
[4] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV, pages 2650–2658, 2015.
[5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, pages 2366–2374, 2014.
[6] R. Garg, V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In ECCV, 2016.
[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In CVPR, 2012.
[8] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In AISTATS, 2010.
[9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. arXiv:1609.03677v1, 2016.
[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In ICCV, 2015.
[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.
[12] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In ECCV, pages 775–788, 2012.
[13] J. Konrad, M. Wang, and P. Ishwar. 2D-to-3D image conversion by learning depth from examples. In CVPR Workshops, 2012.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, NIPS, pages 1097–1105. 2012.
[15] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In CVPR, pages 89–96, 2014.
[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In 3DV, 2016.
[17] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In CVPR, pages 1119–1127, 2015.
[18] C. Li, A. Kowdle, A. Saxena, and T. Chen. Toward holistic scene understanding: Feedback enabled cascaded classification models. PAMI, 34(7):1394–1408, July 2012.
[19] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In CVPR, 2010.
[20] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In CVPR, pages 5162–5170, 2015.
[21] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In CVPR, pages 716–723, 2014.
[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 115(3):211–252, 2015.
[23] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In NIPS, 2005.
[24] A. Saxena, S. H. Chung, and A. Y. Ng. 3D depth reconstruction from a single still image. IJCV, 76(1):53–69. Jan. 2008.
[25] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. PAMI, 2009.
[26] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille. Towards unified depth and semantic prediction from a single image. In CVPR, pages 2800–2809, 2015.
[27] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In ECCV, 2016.