AdaMM-DepthNet: Unsupervised Adaptive Depth Estimation Guided by Min and Max Depth Priors for Monocular Images

Juan Luis Gonzalez Bello, 김문철
한국과학기술원 전기 및 전자 공학과
juanluisgb@kaist.ac.kr, mkim@ee.kaist.ac.kr

AdaMM-DepthNet: Unsupervised Adaptive Depth Estimation Guided by Min and Max Depth Priors for Monocular Images

Juan Luis Gonzalez Bello, Munchurl Kim
Korea Advanced Institute of Science and Technology Dep. Of Electrical Engineering

요 약

Unsupervised deep learning methods have shown impressive results for the challenging monocular depth estimation task, a field of study that has gained attention in recent years. A common approach for this task is to train a deep convolutional neural network (DCNN) via an image synthesis sub-task, where additional views are utilized during training to minimize a photometric reconstruction error. Previous unsupervised depth estimation networks are trained within a fixed depth estimation range, irrespective of its possible range for a given image, leading to suboptimal estimates. To overcome this suboptimal limitation, we first propose an unsupervised adaptive depth estimation method guided by minimum and maximum (min-max) depth priors for a given input image. The incorporation of min-max depth priors can drastically reduce the depth estimation complexity and produce depth estimates with higher accuracy. Moreover, we propose a novel network architecture for adaptive depth estimation, called the AdaMM-DepthNet, which adopts the min-max depth estimation in its front side. Intensive experimental results demonstrate that the adaptive depth estimation can significantly boost up the accuracy with a fewer number of parameters over the conventional approaches with a fixed minimum and maximum depth range.

1. Introduction

Extracting the underlying 3D information of a scene from a single image is the holy grail of computer vision as it opens the door to multiple exciting and useful applications ranging from robotics and navigation to virtual and augmented reality (VR/AR). However, it is now in recent years with the advances in deep learning and convolutional neural networks (CNN) that monocular depth estimation has become a booming research field. The previous classical techniques performed poorly due to the use of fixed assumptions or handcrafted feature extractors. On the other hand, the learning-based approaches automatically learn to extract global and local features useful for depth estimation and can be divided into supervised and unsupervised (or self-supervised) methods. Supervised methods [1, 2] for monocular depth estimation require the hard-to-obtain depth ground truth (GT) data for training and tend to focus on new network architectures. In contrast, the unsupervised methods [3-5], with no depth GT, have an orientation for designing new loss functions or operations that better exploit geometrical constraints between the target view and the additional reference views. These other views usually come in the form of monocular video or stereo pairs, which are only available during training.

2. Method

The motivation for our work arises from observing the behavior of DCNNs trained for monocular depth estimation when configured to estimate depths at different ranges. We first observed that all previous works set fixed min-max depth ranges as hyper-parameters, within which their methods are bound to estimate. In particular, the methods [3-5] that use stereo pairs as training data make predictions in terms of inverse depth or disparity. In these works, the minimum disparity is usually set to 0, and the maximum disparity is set to an arbitrary value smaller than the training sub-image sizes (usually 512 pixels). We trained our implementations of Monodepth [4] and Deep3D [3] with various ranges of inverse depths (or disparities), as shown in Table 1, and noticed that the selection of the minimum and maximum disparity hyper-parameters does affect the performance of the depth estimation networks. To see how the hyper-parameters affect the depth learning, we chose different disparity ranges from 0 to 256, including a minimum disparity of 4.7 pixels, which translates to a depth of roughly 80m in the KITTI [6] dataset, a common maximum depth used for evaluation in most previous works. Interestingly, we observed a tendency of higher accuracy as we reduce the range of the estimated disparities, as shown in Table 1. The effect of minimum and maximum disparity hyper-parameters is more critical for our Deep3D implementation, with a difference of 6.7% in a^1 accuracy between the model trained with a max disparity of 256 and the same model trained with a max disparity of 100. While the numerical results in Table 1 look impressive, and are much better than their original works [3, 4], they should be taken with a grain of salt as the reduced range could increase over-fitting and renders the models useless for closer or farther away objects that lay outside the pre-defined min-max range. For this reason, in conventional approaches, there will always be a trade-off between the performance of depth estimation and the min-max depth range detection when setting these as fixed hyper-parameters. To alleviate this problem, our proposed adaptive depth estimation method provides DCNNs with means of adaptation to per-image depth distributions.

2.1 Proposed minimum and maximum depth estimation

To provide adaptive depth cues to the monocular depth estimation networks, we first pre-train our novel min-max estimation sub-network, the mmd-subnet. Our mmd-subnet is also trained in an unsupervised fashion. For natural scene datasets such as KITTI [6], it is impossible to train the mmd-subnet in a supervised manner, as the provided sparse LiDAR sensor depth GT data does not take into account all objects in the scene and has detection range limitations.

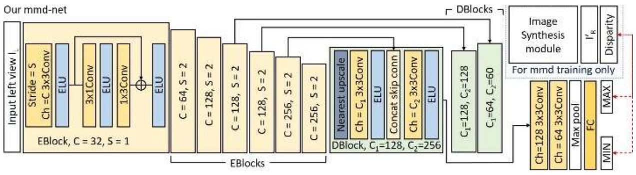

Figure 1. Our min-max depth (mmd) sub-network.

-Mmd-subnet architecture. Figure 1 shows the architecture of our mmd-subnet, which takes as input a single image and outputs two scalars, the minimum and maximum disparity values in the scene. The mmd-subnet has a relatively low parameter count of 4M parameters during inference and performs inference in less than 12ms on a Titan XP GPU on a 375x1242 image. The mmd-subnet has an encoder-decoder-encoder architecture. The first encoder section of the mmd-subnet is made of 7 "EBlocks", each of which is made of one convolution and one residual block. To reduce the parameter count and inference time, we utilize 3x1 and 1x3 convolutions in the residual blocks. The decoder part is made of 3 "DBlocks", each of which is made of a nearest upscale operation followed by a convolution, a skip connection from the encoder part, and another conv layer. During testing, only one DBlock is needed to generate the features for min-max depth estimation. These features are encoded again into a fewer channel feature map by two conv layers and then reduced to a vector via global max pooling. We observed that global max-pooling yielded better results than global average pooling, which is understandable, as the min-max depth priors are in highly localized pixel locations and cannot be well represented by an averaging operation. A fully connected layer is then incorporated for the final estimation of the min-max depths, d_{min} and d_{max}, of the single image input.

Table 1. Training Monodepth [4] and Deep3D [3] with different min-max disparity ranges (pixel units) the KITTI2015 train dataset [6]. Arrows indicate the better metric

Min/Max	abs rel↓	sq rel↓	rms↓	Log rmse↓	a ↑
Monodepth [4]					
0/256	0.136	2.513	6.256	0.277	0.878
0/154	**0.132**	**2.421**	**6.165**	0.226	0.880
4.7/135	0.135	2.445	6.180	**0.224**	**0.881**
4.7/100	0.134	2.511	6.241	0.227	0.880
Deep3D [3]					
0/256	0.134	1.206	6.483	0.242	0.830
0/154	0.140	1.197	6.262	0.231	0.831
4.7/135	0.103	0.832	5.431	0.185	0.881
4.7/100	**0.094**	**0.710**	**4.717**	**0.165**	**0.897**

-Training strategy of the mdd-subnet. During training only, the decoder part gradually upscales and fuses the encoder features up to a scale of 1/8 of the input image, where the resulting features are nearest-upscaled to the input resolution and fed to our image synthesis module (explained in depth in Section 2.2) for right-view synthesis of I_R^{mm} and the disparity estimation of D'_{mm} . This module aids in the training of the first encoder-decoder and provides min-max proxy labels for the second encoder part. We adopted our image synthesis module not only because it showed state-of-the-art results for depth estimation, but more importantly, because it presents considerably good detection of thin objects and complex structures, critical for min-max depth prediction. The mmd-subnet is trained to minimize a min-max depth loss and a synthesis loss between the synthetic image I_R^{mm} and the GT right view I_R, given by

$$l_{mmd} = l_1\big(d_{min}, min(D'_{mm})\big) + l_1\big(d_{max}, max(D'_{mm})\big) + l_{syn}(I_R^{mm}, I_R) \quad (1)$$

2.2 Adaptive depth guided by min-max depth priors

To perform adaptive depth estimation, a DCNN first needs the minimum and maximum depth information for any given single-view input image. Any DCNN for monocular depth estimation can perform adaptive depth estimation by attaching our novel mmd-subnet upfront. The min and max depth values are stretched to the input image resolution and concatenated with the input view to provide prior knowledge to the network. Similarly, the min-max depth values are used at the network's output to re-scale the final depth predictions for the networks that perform direct estimation, as in [4], and to control the sampling positions or receptive field sizes of the adaptive convolutional operations for the networks that obtain depth with indirect methods, as in [3]. A ±5% is allowed in the estimated min-max depth priors to handle the possible inaccuracies in the mmd-subnet.

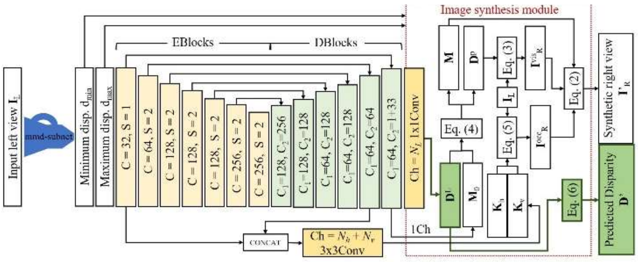

Figure 2. Our proposed AdaMM-DepthNet. Operations needed for depth estimation are shown filled in green.

2.3 Our proposed AdaMM-DepthNet

With the min-max depth prior estimated by our mmd-subnet, we construct an adaptive monocular depth estimation network, the AdaMM-DepthNet (depicted in Figure 2), which estimates full-resolution depth maps while maintaining a lower number of parameters, compared to the previous works [2-5]. We train our AdaMM-DepthNet via a stereoscopic view synthesis sub-task from a single left-view image I_L for which we define the following image synthesis model

$$I'_R = (1 - M) \odot I_R^{vis} + M \odot I_R^{occ} \qquad (2)$$

which describes a synthetic right-view image I'_R as a combination of the visible (in both views) right image contents I_R^{vis} and the occluded contents I_R^{occ} which are only visible in the right view. $\odot$ denotes the element-wise multiplication operation. M is the blending mask between the visible and occluded pixels. Due to the geometrical constraints in stereo image pairs, the I_R^{vis} is highly correlated to the depth (or disparity) of the scene. Note that some previous methods ignore occlusions in their formation models [5] and others include occlusions in a single operation [3]. In contrast, our synthesis model explicitly generates both visible and occluded content images and blends them to generate the final synthetic right-view. The visible contents image is defined as

$$I_R^{vis} = \sum_{i=0}^{N_L} g_h\left(I_L, \frac{i}{N_L}(d_{max} - d_{min}) + d_{min}\right) \odot D_i^p \qquad (3)$$

where $g_{h(v)}(\cdot)$ is a horizontal (vertical) shift operation, which can take fractional units (supported by bilinear interpolation), N_L is the number planes, and D^p is the disparity probability volume, which is given by

$$M, D^p = \sigma\left(M_0, \left\{g_h\left(D_i^L, \frac{i}{N_L}(d_{max} - d_{min}) + d_{min}\right)\right\}_{i=0}^{N_L}\right) \qquad (4)$$

where D^L are the disparity logits generated by the AdaMM-DepthNet. M_0 is an output of our network that helps guiding the channel-wise softmax operation $\sigma(\cdot)$ when obtaining M, D^p. To estimate the occluded content image I_R^{occ}, we make use of our implementation of adaptive separable convolutions by

$$I_R^{occ} = \sum_{i=0}^{N_v} g_v\left(\sum_{j=0}^{N_h} g_h\left(I_L, -\frac{0.5j}{N_h} d_{max}\right) \odot K_h, \frac{i - 0.5N_v}{N_v} d_{max}\right) \odot K_v \qquad (5)$$

where K_h and K_v are the horizontal and vertical 1D kernel components generated by our AdaMM-DepthNet with N_h and N_v number of kernel elements respectively. It can be noted in the use of the negative sign in the horizontal component operation that Eq. (5) only samples pixels to the opposite side (including upper and lower pixels) to the operation in Eq. (3). This is done intentionally in order to enforce the network to keep stereo correspondences on Eq. (3), in other words, K_h and K_v prevent the disparity logits D^L from learning occlusion information. Under this condition, the final disparity estimate D' can be obtained from D^L by

$$D' = d_{min} + (d_{max} - d_{min}) \sum_{i=1}^{N_L} \frac{i}{N_L} \sigma(D^L)_i \qquad (6)$$

2.4 Unsupervised loss functions for depth estimation

To train the Monodepth for adaptive depth estimation, the photometric, smoothness, and consistency loss functions defined in its original work [4] were utilized. For training our Deep3D implementation and our AdaMM-DepthNet for adaptive depth estimation, we use an image synthesis loss which is a combination of l_1 and perceptual loss l_p. l_p is balanced by α_p, which was empirically set to 0.01, as given by

$$l_{syn} = l_1 + \alpha_p l_p = |I_R - I'_R|_1 + \alpha_p \sum_{l=1}^{3} \left\|\phi^l(I_R) - \phi^l(I'_R)\right\|_2^2 \qquad (7)$$

3. Experiments

3.1 Details of implementation, datasets and evaluations

For fair and extensive comparison with previous works, we train all our networks for 50 epochs on the KITTI Eigen train split [1] by the Adam optimizer, with an initial learning rate of 0.0001 (halved at epochs 30 and 40), a batch size of 8 and a sub-image size of 256x512. During training, data augment-ations are incorporated on the fly with random crop, random horizontal flip, random gamma, random brightness, and individual color brightness. The KITTI Eigen split [1] consists of 22,600 image pairs that are selected from the stereo KITTI [6] dataset, avoiding static car frames to ensure diversity. To measure the performance of our networks, we use the various KITTI metrics defined in [1]. We ablate our networks on the KITTI2015 scene-flow training dataset [6], which contains 200 images with CAD-refined sparse LiDAR depth ground truth. To intensively compare ours against a broader spectrum of previous works, we test our method on the improved KITTI Eigen test split [1], which contains 652 images with sparse (but improved) LiDAR depth GT of scenes excluded from the training split.

3.2 Ablation studies

To show the effectiveness of our proposed adaptive depth estimation, we compared the conventional methods, Monodepth [4] and Deep3D [3], against their variations with our mmd-subnet upfront which are called Monodepth-mmd and Deep3D-mmd, respectively. In Table 2, it is obvious that the two variations (Monodepth-mmd and Deep3D-mmd), significantly outperform the conventional methods with fixed min-max disparity hyperparameters. In particular, the Deep3D-mmd, showed a *significant* improvement of 7.3% in a^1 accuracy versus its conventional counterpart.

3.3 Results on KITTI

Table 3 compares our AdaMM-DepthNet against various existing works for unsupervised and supervised depth estimation on the KITTI Eigen test split [1]. As can be noted, our AdaMM-DepthNet *remarkably* outperforms all the unsupervised methods, even without any post-processing (PP) step nor stereo-SMG supervision, which are required in the previous SOTA methods of DepthHints [5]. When applying a PP step following [4], our AdaMM-DepthNet outperforms the previous supervised method of DORN [2] in most metrics. Additionally, our AdaMM-DepthNet achieves superior performance with a modest number of parameters (22M), in comparison with the previous SOTA (35M in DepthHints[5] and 51M in DORN [2]). Figure 3 shows the visual comparison of our method against the previous works. It can be observed that our AdaMM-DepthNet produces more consistent depths with better detection of fine details in thin structures like traffic signs and pedestrians.

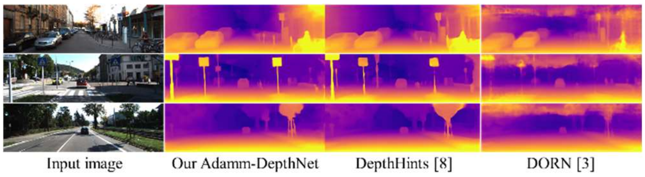

Figure 3. Qualitative comparison with other methods.

Table 2. Effect of min-max depth priors on Monodepth [4] and Deep3D [3] on the KITTI2015 dataset [6].

Network	abs rel↓	sq rel↓	RMSE↓	$\mathrm{RMSE}_{\log}$↓	a ↑
Monodepth [4]	0.136	2.513	6.265	0.227	0.878
Monodepth-mmd	0.135	2.435	6.173	0.228	0.880
Deep3D [3]	0.134	1.206	6.483	0.242	0.830
Deep3D-mmd	**0.096**	**0.671**	**4.403**	**0.159**	**0.903**

4. Conclusions

We presented a novel unsupervised adaptive depth estimation method for monocular images, which incorporates our novel mmd-subnet for min-max depth prior estimation at the front. Through extensive experiments, we showed that such min-max depth priors are fundamentally important for the depth estimation task, which enables our monocular depth estimation network, the AdaMM-DepthNet to outperform the recent SOTA methods. Furthermore, we proposed an effective image synthesis model that can generate occluded and visible components in synthesized images, allowing for better learning of monocular image depth. This image synthesis model is used to train our proposed AdaMM-DepthNet, which also incorporates our mmd-subnet upfront and outperforms previous self-supervised methods that learn from stereo images by considerable margins.

Table 3. Evaluation on the KITTI Improved Eigen split [1]. PP: post-processing. Our method achieves the best performance.

Network	abs rel↓	sq rel↓	RMSE↓	$\mathrm{RMSE}_{\log}$↓	a ↑
DORN [2]	0.072	0.307	**2.727**	0.120	0.932
DepthHints [5] (PP)	0.074	0.364	3.202	0.114	0.936
AdaMM-DepthNet	0.073	0.317	2.995	0.110	0.938
AdaMM-DepthNet (PP)	**0.070**	**0.300**	2.906	**0.108**	**0.940**

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00419, Intelligent High Realistic Visual Processing for Smart Broadcasting Media).

References

[1] Eigen, D., Puhrsch, C., Fergus, R.: “Depth map prediction from a single image using a multi-scale deep network”. In: Advances in neural information processing systems (2014).

[2] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: “Deep ordinal regression network for monocular depth estimation”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018).

[3] Xie, J., Girshick, R., Farhadi, A.: “Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks”. In: European Conference on Computer Vision (2016).

[4] Godard, C., Mac Aodha, O., Brostow, G.J.: “Unsupervised monocular depth estimation with left-right consistency”. In: The IEEE Conference on Computer Vision and Pattern Recognition (2017).

[5] Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: “Self-supervised monocular depth hints”. In: The IEEE International Conference on Computer Vision (2019).

[6] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (2012).