

From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation

Jin Han Lee, Myung-Kyu Han, Dong Wook Ko and Il Hong Suh
 Department of Electronics and Computer Engineering, Hanyang University
 {jinhanlee, mkhan91, pumpblack, ihsuh}@hanyang.ac.kr

Abstract

Estimating accurate depth from a single image is challenging because it is an ill-posed problem as infinitely many 3D scenes can be projected to the same 2D scene. However, recent works based on deep convolutional neural networks show great progress with plausible results. The convolutional neural networks are generally composed of two parts: an encoder for dense feature extraction and a decoder for predicting the desired depth. In the encoder-decoder schemes, repeated strided convolution and spatial pooling layers lower the spatial resolution of transitional outputs, and several techniques such as skip connections or multi-layer deconvolutional networks are adopted to recover the original resolution for effective dense prediction.

In this paper, for more effective guidance of densely encoded features to the desired depth prediction, we propose a network architecture that utilizes **novel local planar guidance layers located at multiple stages in the decoding phase**. We show that the proposed method outperforms the state-of-the-art works with significant margin evaluating on challenging benchmarks. We also provide results from an ablation study to validate the effectiveness of the proposed method.

1. Introduction

Depth estimation from 2D images has been studied in computer vision for a long time and is nowadays applied to robotics, autonomous driving cars, scene understanding, and 3D reconstruction. Those applications usually utilize, to perform depth estimation, multiple instances of the same scene such as stereo image pairs [40], multiple frames from moving camera [34] or static captures under different lighting conditions [1, 2]. As depth estimation from multiple observations achieves impressive progress, it naturally leads to depth estimation with a single image since it demands less cost and constraint.

However, estimating accurate depth from a single im-

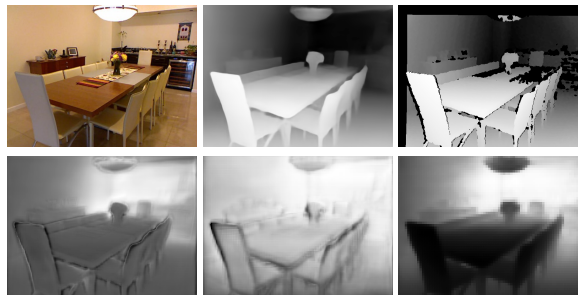


Figure 1: Example outputs from the proposed network. Top: from left to right, input image, predicted depth map, and the ground truth. Bottom: from left to right, outputs from the proposed local planar guidance layers having input feature resolutions of 1/2, 1/4, 1/8 to the input image, respectively.

age is challenging, even for a human, because it is an ill-posed problem as infinitely many 3D scenes can project to the same 2D scene. To understand geometric configuration from a single image, humans consider use not only local cues, such as texture appearance in various lighting and occlusion conditions, perspective, or relative scales to the known objects, but also **global context, such as entire shape or layout of the scene** [19].

After the first learning-based monocular depth estimation work from Saxena et al. [38] was introduced, considerable improvements have been made along with rapid advances in deep learning [11, 10, 28, 29, 45, 35, 21, 25]. While most of the state-of-the-art works apply models based on deep convolutional neural networks (DCNNs) in *supervised* fashion, some works proposed *semi-* [24] or *self-supervised* learning methods which do not entirely rely on the ground truth depth data.

In the meantime, recent applications based on DCNNs are commonly composed in two parts: encoder for dense feature extraction and decoder for the desired prediction. **As a dense feature extractor, very powerful deep networks such as VGG [43], ResNet [18] or DenseNet [20] are usu-**

ally adopted. In these networks, repeated strided convolution and spatial pooling layers lower the spatial resolution of transitional outputs, which can be a bottleneck to obtain desired dense predictions. Therefore, a number of techniques, for example, multi-scale networks [29, 10], skip connections [16, 47] or multi-layer deconvolutional networks [25, 14, 24] are applied to consolidate feature maps from higher resolutions. Recently, atrous spatial pyramid pooling (ASPP) [5] has been introduced for image semantic segmentation, which can capture large scale variations in observation by applying sparse convolutions with various dilation rates. Since the dilated convolution allows larger receptive field size, recent works in semantic segmentation [5, 50] or depth estimation [12] do not fully reduce the receptive field size by removing last few pooling layers and reconfigure the network with atrous convolutions to reuse pre-trained weights. Consequently, their methods have larger dense features (1/8 of input spatial resolution whereas 1/32 or 1/64 in the original base networks) and perform almost all of the decoding process on that resolution followed by a simple upsampling to recover the original spatial resolution.

To define explicit relation in recovering back to the full resolution, we propose a network architecture that utilizes novel local planar guidance layers located at multiple stages in the decoding phase. Specifically, based on an encoding-decoding scheme, at each decoding stage, which has spatial resolutions of 1/8, 1/4, and 1/2, we place a layer that effectively guides input feature maps to the desired depth with local planar assumption. Then, we combine the outputs to predict depth in full resolution. This differs from multi-scale network [10, 11] or image pyramid [16] approaches in two aspects. First, the outputs from the proposed layers are not treated as separated global depth estimation in corresponding resolutions. Instead, we let the layers to learn 4-dimensional plane coefficients and use them together to reconstruct depth estimations in the full resolution for the final estimation. Second, as a consequence of the nonlinear combination, individual spatial cells in each resolution are distinctively trained while the training progresses. We can see example outputs from the proposed layers in Figures 1 and 3. Experiments on the challenging NYU Depth V2 dataset [42] and KITTI dataset [15] demonstrate that the proposed method achieves state-of-the-art results.

The rest of the paper is organized as follows. After a concise survey of related works in Section 2, we present in detail the proposed method in Section 3. Then, in Section 4, we provide results on two challenging benchmarks comparing with state-of-the-art works, and using various base networks as an encoder for the proposed network, we see how the performance varies along with each base network. In Section 4, we also provide an ablation study to validate the effectiveness of the proposed method. We conclude the

paper in Section 5.

2. Related Work

2.1. Supervised Monocular Depth Estimation

In monocular depth estimation, supervised approaches take a single image and use depth data measured with range sensors such as RGB-D cameras or multi-channel laser scanners as ground truth for supervision in training. Saxena et al. [38] propose a learning-based approach to get a functional mapping from visual cues to depth via Markov random field, and extend it to a patch-based model that first over-segments the input image and learns 3D orientation as well as the location of local planes that are well explained by each patch [39]. Eigen et al. [10] introduce a multi-scale convolutional architecture that learns coarse global depth predictions on one network and progressively refine them using another network. Unlike the previous works in single image depth estimation, their network can learn representations from raw pixels without handcrafted features such as contours, super-pixels, or low-level segmentation. Several works follow the success of this approach by incorporating strong scene priors for surface normal estimation [46], using conditional random fields to improve accuracy [27, 23, 41] or changing the learning problem from regression to classification [3]. A recent supervised approach from Fu et al. [12] achieves the state-of-the-art result by also taking advantage of changing the regression problem to quantized ordinal regression. Xu et al. [49] propose an architecture that exploits multi-scale estimations derived from inner layers by fusing them within a CRF framework. Gan et al. [13] propose to explicitly model the relationships between different image locations with an affinity layer. Most recently, Yin et al. [51] introduce a method using virtual normal directions that are determined by randomly chosen three points in the reconstructed 3D space as geometric constraints.

2.2. Semi-Supervised Monocular Depth Estimation

There are also attempts to train a depth estimation network in a semi-supervised or weakly supervised fashion. Chen et al. [6] propose a new approach that uses information of relative depth and depth ranking loss function to learn depth predictions in unconstrained images. Recently, to overcome the difficulty in getting high-quality depth data, Kuznetsov et al. [24] introduce a semi-supervised method to train the network using both sparse LiDAR depth data for direct supervision and image alignment loss as a secondary training objective.

2.3. Self-Supervised Monocular Depth Estimation

The self-supervised approach refers to a method that requires only rectified stereo image pairs to train the depth

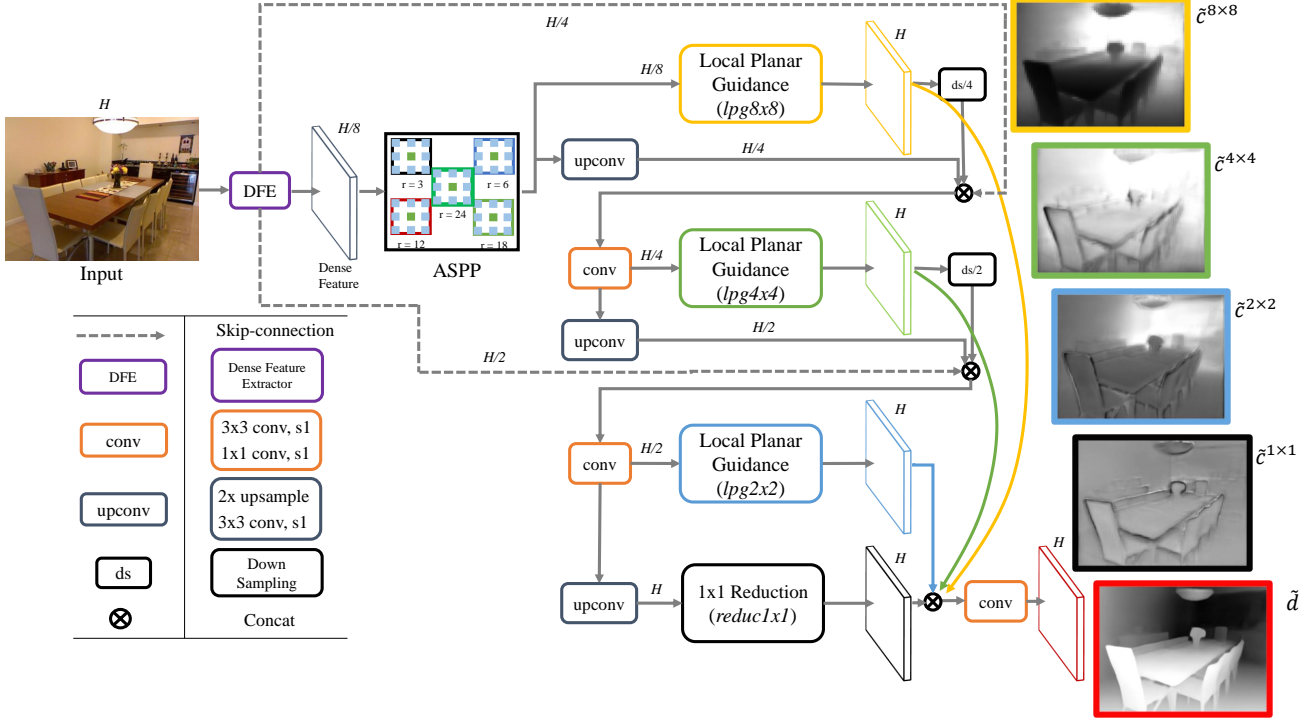


Figure 2: Overview of the proposed network architecture. The network is composed of dense feature extractor (the base network), contextual information extractor (ASPP), local planar guidance layers and their dense connection for final depth estimation. Note that the outputs from the local planar guidance layers have the full spatial resolution H enabling shortcuts inside the decoding phase. We also use skip-connections from the base network to link with internal outputs in the decoding phase with corresponding spatial resolutions.

estimation network. Garg et al. [14] and Godard et al. [16] propose *self-supervised* learning methods that smartly cast the problem from direct depth estimation to image reconstruction. Specifically, with a rectified stereo image pair, their networks try to synthesize one view from the other with estimated disparities and define the error between both as the reconstruction loss for the main training objective. In this way, because learning requires only well rectified, synchronized stereo pairs instead of the ground truth depth data well associated with the corresponding RGB images, it greatly reduces the effort to acquire datasets for new categories of scenes or environments. However, there is some accuracy gap when compared to the current best supervised approach [51]. Garg et al. [14] introduce an encoder-decoder architecture and to train the network using photometric reconstruction error. Xie et al. [47] propose a network that also synthesizes one view from the other, and by using the reconstruction error, they produce a probability distribution of possible disparities for each pixel. Godard et al. [16] finally propose a network architecture fully differentiable thus can perform end-to-end training. They also present a novel left-right consistency loss that improves

training and predictions of the network. Most recently, Godard et al. [17] propose a simple but effective architecture benefiting from associated design choices such as a robust reprojection loss, multi-scale sampling, and an auto-masking loss.

2.4. Video-Based Monocular Depth Estimation

There are also approaches using sequential data to perform the monocular depth estimation. Yin et al. [52] propose an architecture consisting of two generative sub-networks that are jointly trained by adversarial learning for disparity map estimation organized in a cycle to provide mutual constraints. Mahjourian et al. [30] presents an approach that explicitly considers the inferred 3D geometry of the whole scene, and enforce consistency of the estimated 3D point clouds and ego-motion across consecutive frames. Wang et al. [44] adopt a differentiable pose predictor and train a monocular depth estimation network in an end-to-end fashion while benefiting from the pose predictor.

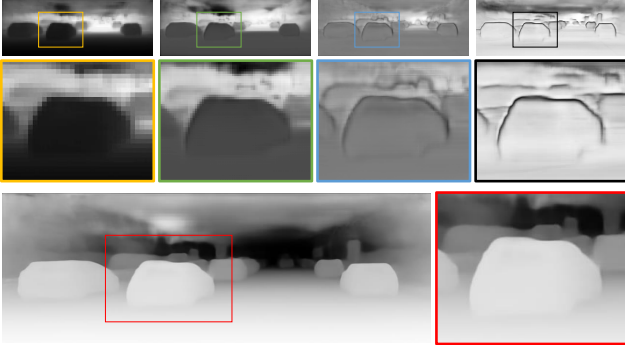


Figure 3: Examples showing behavior of the proposed network. Top and middle rows show $\tilde{c}^{k \times k}$ and their focused views. Bottom row shows the final depth estimation result with a focused view. The overestimated boundaries of the vehicle from lpg8x8 (yellow rectangle) and lpg4x4 (green rectangle) are compensated by the outputs from lpg2x2 (blue rectangle) and reduclx1 (black rectangle) (i.e., $\tilde{c}^{2 \times 2}$ and $\tilde{c}^{1 \times 1}$), resulting in the clear boundary in the final estimation.

3. Method

In this section, we describe the proposed monocular depth estimation network with a novel local planar guidance layer located on multiple stages in the decoding phase.

3.1. Network Architecture

As it can be seen from Figure 2, we follow an encoding-decoding scheme that reduces feature map resolution to $H/8$ then recovers the original resolution H for dense prediction. After the backbone network that we use as a dense feature extractor which produces an $H/8$ feature map, we place a denser version [50] of atrous spatial pyramid pooling layer [5] as our contextual information extractor with various dilation rates $r \in \{3, 6, 12, 18, 24\}$. Then, at each stage in the decoding phase, where internal outputs are recovered to the full spatial resolution with a factor of 2, we employ the proposed local planar guidance (LPG) layer to locate geometric guidance to the desired depth estimation. We also place a 1×1 reduction layer to get the finest estimation $\tilde{c}^{1 \times 1} \in R^{H \times W \times 1}$ after the last *upconv* layer. Finally, outputs from the proposed layers (i.e., $\tilde{c}^{k \times k}$) and $\tilde{c}^{1 \times 1}$ are concatenated and fed into the final convolutional layer to get the depth estimation \tilde{d} .

3.2. Multi-Scale Local Planar Guidance

Our key idea in this work is to define direct and explicit relations between internal features and the final output in an effective manner. Unlike the existing methods that recover back to the original resolution using simple nearest neighbor upsampling layers and skip connections from encoding stages, we place novel local planar guidance layers which

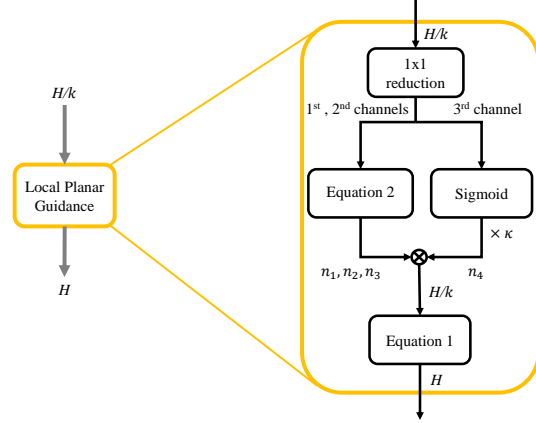


Figure 4: The local planar guidance layer. We use a stack of 1×1 convolutions to get 4D coefficients estimations. (i.e., $H/k \times H/k \times 4$). Then the channels are split to pass through two different activation mechanisms to ensure plane coefficients' constraint. Finally, they are fed into the planar guidance module to compute locally-defined relative depth estimations.

guide features to the full resolution with the local planar assumption and use them together to get the final depth estimation \tilde{d} . As can be seen from Figure 2, since the proposed layer recovers given an internal feature map to the full resolution H , it can be used as a skip connection inside the decoding phase allowing direct relations between internal features and the final prediction. Specifically, given a feature map having spatial resolution H/k , the proposed layers estimate for each spatial cell a 4D plane coefficients that fit a locally defined $k \times k$ patch on the full resolution H , and they are concatenated together for the prediction through the final convolutional layers.

Please note that the proposed LPG layer is not designed to directly estimate global depth values on the corresponding scale because the training loss is only defined in terms of the final depth estimation (provided in Section 3.3). Together with outputs from the other LPG layers and *reduclx1*, each output is interpreted to the global depth by contributing as a part of the nonlinear combination through the final convolutional layers. Therefore, they can have distinct ranges, learned as a base or precise relative compensation from the base at a spatial location, as shown in Figures 1 and 3.

Here, we use the local planar assumption because, for a $k \times k$ region, it enables an efficient reconstruction with only four parameters. If we adopt typical *upconvs* for the reconstruction, the layers should be learned to have k^2 values properly instead of four. Therefore, we can expect that our strategy can be more effective because conventional upsampling would not give details on enlarged resolutions, while

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	AbsRel	RMSE	log10
Saxena et al. [39]	0.447	0.745	0.897	0.349	1.214	-
Wang et al. [45]	0.605	0.890	0.970	0.220	0.824	-
Liu et al. [29]	0.650	0.906	0.976	0.213	0.759	0.087
Eigen et al. [10]	0.769	0.950	0.988	0.158	0.641	-
Chakrabarti et al. [4]	0.806	0.958	0.987	0.149	0.620	-
Li et al. [28]	0.789	0.955	0.988	0.152	0.611	0.064
Laina et al. [25]	0.811	0.953	0.988	0.127	0.573	0.055
Xu et al. [49]	0.811	0.954	0.987	0.121	0.586	0.052
Lee et al. [26]	0.815	0.963	0.991	0.139	0.572	-
Fu et al. [12]	0.828	0.965	0.992	0.115	0.509	0.051
Qi et al. [33]	0.834	0.960	0.990	0.128	0.569	0.057
Yin et al. [51]	0.875	0.976	0.994	0.108	0.416	0.048
Ours-ResNet	0.871	0.977	0.995	0.113	0.407	0.049
Ours-ResNext	0.880	0.977	0.994	0.111	0.410	0.048
Ours-DenseNet	0.885	0.978	0.994	0.110	0.392	0.047

Table 1: Evaluation results on NYU Depth v2. Ours outperforms previous works with a significant margin in all measures except only from AbsRel.

the local linear assumption can provide effective guidance.

To guide features with the local planar assumption, we convert each estimated 4D plane coefficients to $k \times k$ local depth cues using *ray-plane intersection*:

$$\tilde{c}_i = \frac{n_4}{n_1 u_i + n_2 v_i + n_3}, \quad (1)$$

where $n = (n_1, n_2, n_3, n_4)$ denotes the estimated plane coefficients, (u_i, v_i) are $k \times k$ patch-wise normalized coordinates of pixel i .

Figure 4 shows the detail of the proposed layer. Through a stack of 1×1 convolutions which repeatedly reduce the number of channels by a factor of 2 until it reaches 3, we get a $H/k \times H/k \times 3$ feature map if we assume a square input. Then, we pass the feature map through two different ways to get local plane coefficient estimations: one way is a conversion to a unit normal vector (n_1, n_2, n_3) , and the other is a sigmoid function defining the perpendicular distance n_4 between the plane and origin. After the sigmoid function we multiply the output with the maximum distance κ to get real depth values. Because a unit normal vector has only two degrees of freedom (i.e., polar and azimuthal angles θ, ϕ from predefined axes), we regard the first two channels of the given feature map as the angles and convert them to unit normal vectors using following equations.

$$n_1 = \sin(\theta) \cos(\phi), n_2 = \sin(\theta) \sin(\phi), n_3 = \cos(\theta). \quad (2)$$

Finally, they are concatenated again and used for estimation of $\tilde{c}^{k \times k}$ using Equation 1.

We design the local depth cue as an additive depth defined in local regions (i.e., $k \times k$ patches). Since features at the same spatial location in different stages are used together to predict the final depth, for efficient representation, we expect that global shapes would be learned at coarser scales while local details at finer scales. Also, they can interact with each other to compensate for erroneous estimations. We can represent the behavior of the last convolu-

tional layer as follows.

$$\tilde{d} = f(W_1 \tilde{c}^{1 \times 1} + W_2 \tilde{c}^{2 \times 2} + W_3 \tilde{c}^{4 \times 4} + W_4 \tilde{c}^{8 \times 8}), \quad (3)$$

where f is an activation function, $W_j, j \in \{1, 2, 3, 4\}$ denotes a corresponding linear transform representing the convolution. Please note that the proposed network learns on multiple scales, and by defining the training loss only in terms of the final estimation, \tilde{d} , we do not enforce parameters for each scale learns with the constant contribution. Therefore, in training, details for regions with sharp curvatures would be learned at finer scales while major structures at coarser scales. Also, there is the last chance in $\tilde{c}^{1 \times 1}$ to recover broken assumptions in the upsampled estimations ($\tilde{c}^{k \times k}, k \in \{2, 4, 8\}$). From Figure 3, we can see small details behind the focused vehicle from blue- and black-boxed figures which demonstrate $\tilde{c}^{2 \times 2}$ and $\tilde{c}^{1 \times 1}$, respectively, while they are missing in the coarser scales, $\tilde{c}^{8 \times 8}$ and $\tilde{c}^{4 \times 4}$. Also, there are thick black estimations in $\tilde{c}^{1 \times 1}$ and $\tilde{c}^{2 \times 2}$ on the boundary of the vehicle compensating the over-estimations in $\tilde{c}^{8 \times 8}$ and $\tilde{c}^{4 \times 4}$. More examples are provided in the supplementary material.

3.3. Training Loss

In [11], Eigen et al introduce a scale-invariant error and inspired from it they use a following training loss:

$$D(g) = \frac{1}{T} \sum_i g_i^2 - \frac{\lambda}{T^2} \left(\sum_i g_i \right)^2, \quad (4)$$

where $g_i = \log \tilde{d}_i - \log d_i$ with the ground truth depth d_i , $\lambda = 0.5$ and T denotes the number of pixels having valid ground truth values. By rewriting above equation,

$$D(g) = \frac{1}{T} \sum_i g_i^2 - \left(\frac{1}{T} \sum_i g_i \right)^2 + (1 - \lambda) \left(\frac{1}{T} \sum_i g_i \right)^2,$$

we can see that it is a sum of the variance and a weighted squared mean of the error in log space. Therefore, setting a higher λ enforces more focusing on minimizing the variance of the error, and we use $\lambda = 0.85$ in this work. Also, we observe that properly scaling the range of the loss function improves convergence and the final training result. Finally, we define our training loss L as follows:

$$L = \alpha \sqrt{D(g)}, \quad (5)$$

where α is a constant we set to 10 for all experiments.

4. Experiments

To verify the effectiveness of our approach, we provide experimental results from challenging benchmarks with

Method	cap	<i>higher is better</i>			<i>lower is better</i>			
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE log
Saxena et al. [39]	0-80m	0.601	0.820	0.926	0.280	3.012	8.734	0.361
Eigen et al. [11]	0-80m	0.702	0.898	0.967	0.203	1.548	6.307	0.282
Liu et al. [29]	0-80m	0.680	0.898	0.967	0.201	1.584	6.471	0.273
Godard et al. (CS+K) [16]	0-80m	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Kuznetsov et al. [24]	0-80m	0.862	0.960	0.986	0.113	0.741	4.621	0.189
Godard et al. (CS+K) [16]	0-80m	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Gan et al. [13]	0-80m	0.890	0.964	0.985	0.098	0.666	3.933	0.173
Fu et al. [12]	0-80m	0.932	0.984	0.994	0.072	0.307	2.727	0.120
Yin et al. [51]	0-80m	0.938	0.990	0.998	0.072	-	3.258	0.117
Ours-ResNet	0-80m	0.954	0.992	0.998	0.061	0.261	2.834	0.099
Ours-DenseNet	0-80m	0.955	0.993	0.998	0.060	0.249	2.798	0.096
Ours-ResNext	0-80m	0.956	0.993	0.998	0.059	0.245	2.756	0.096
Garg et al. [14]	0-50m	0.740	0.904	0.962	0.169	1.080	5.104	0.273
Godard et al. (CS+K) [16]	0-50m	0.873	0.954	0.979	0.108	0.657	3.729	0.194
Kuznetsov et al. [24]	0-50m	0.875	0.964	0.988	0.108	0.595	3.518	0.179
Gan et al. [13]	0-50m	0.898	0.967	0.986	0.094	0.552	3.133	0.165
Fu et al. [12]	0-50m	0.936	0.985	0.995	0.071	0.268	2.271	0.116
Ours-ResNet	0-50m	0.962	0.994	0.999	0.058	0.183	1.995	0.090
Ours-DenseNet	0-50m	0.964	0.995	0.999	0.057	0.175	1.949	0.088
Ours-ResNext	0-50m	0.964	0.994	0.999	0.056	0.169	1.925	0.087

Table 2: Performance on KITTI Eigen split. (CS+K) denotes a model pre-trained on Cityscapes dataset [8] and fine-tuned on KITTI.

Method	SILog	sqErrorRel	absErrorRel	iRMSE
Yin et al. [51]	12.65	2.46	10.15	13.02
Diaz et al. [9]	12.39	2.49	10.10	13.48
Fu et al. [12]	11.77	2.23	8.78	12.98
Ours	11.67	2.21	9.04	12.23

Table 3: Result on the online KITTI evaluation server.

various settings. After presenting the implementation details of our method, we provide experimental results on two challenging benchmarks covering both indoor and outdoor environments. We also provide scores on the online KITTI evaluation server comparing with published works. Then, we provide an ablation study to discuss a detailed analysis of the proposed core factors, and some qualitative results to demonstrate our approach comparing with competitors.

4.1. Implementation Details

We implement the proposed network using the open deep learning framework *PyTorch* [32]. For training, we use Adam optimizer [22] with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-6}$, learning is scheduled via polynomial decay from base learning rate 10^{-4} with power $p = 0.9$. The total number of epochs is set to 50 with batch size 16 on a desktop equipped with four NVIDIA 1080ti GPUs for all experiments in this work.

As the backbone network for dense feature extraction, we use ResNet-101 [18], ResNext-101 [48] and DenseNet-161 [20] with pretrained weights trained for image classification using ILSVRC dataset [36]. Because weights at early convolutions are known to be well trained for primitive visual features, in the base networks, we fix the first two convolutional layers as well as batch normalization parameters in our training. Following [16], we use exponential linear units [7] as an activation function, and *upconv* uses the nearest neighbor upsampling followed by a 3×3 convolution layer [31].

To avoid over-fitting, we augment images before input to the network using random horizontal flipping as well as random contrast, brightness, and color adjustment in a range of $[0.9, 1.1]$, with 50% of chance. We also use a random rotation of the input images in ranges of $[-1, 1]$ and $[-2.5, 2.5]$ degrees for KITTI and NYU datasets, respectively. We train our network on a random crop of size 352×704 for KITTI and 416×544 for NYU Depth V2 datasets.

4.2. NYU Depth V2 Dataset

The NYU Depth V2 dataset [42] contains 120K RGB and depth pairs having a size of 480×640 acquired as video sequences using a Microsoft Kinect from 464 indoor scenes. We follow the official train/test split as previous works, using 249 scenes for training and 215 scenes (654 images) for testing. From the total 120K image-depth pairs,

Variant	# Params	<i>higher is better</i>			<i>lower is better</i>				
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE log	log10
Baseline	63.0M	0.815	0.958	0.990	0.142	0.084	0.587	0.169	0.053
Baseline + A	67.4M	0.827	0.964	0.992	0.141	0.081	0.577	0.166	0.051
Baseline + A + U	68.4M	0.845	0.967	0.992	0.134	0.076	0.513	0.161	0.051
Baseline + A + U + L	68.5M	0.863	0.974	0.994	0.119	0.072	0.421	0.149	0.050
Ours-ResNet	68.5M	0.871	0.975	0.995	0.113	0.068	0.407	0.148	0.049
Ours-DenseNet	47.0M	0.885	0.978	0.994	0.110	0.066	0.392	0.142	0.047

Table 4: Result from the ablation study using the NYU Depth V2 dataset. Baseline: a network composed of only the dense feature extractor and direct estimation from it followed by an upsampling with a factor of 8, A: ASPP module attached after the dense feature extractor, U: using *upconv* layers in Figure 2, L: the proposed local planar guidance layers. All variants are trained using *ResNet-101* as the base network and Equation 4 with $\lambda = 0.5$ as the training loss. ‘Ours-ResNet’ and ‘Ours-DenseNet’ use the training loss given in Equation 5.

due to asynchronous capturing rates between RGB images and depth maps, we associate and sample them using timestamps by even-spacing in time, resulting in 24231 image-depth pairs for the training set. Using raw depth images and camera projections provided by the dataset, we align the image-depth pairs for accurate pixel registrations. We use $\kappa = 10$ for this dataset.

4.3. KITTI Dataset

KITTI provides the dataset [15] with 61 scenes from “city”, “residential”, “road” and “campus” categories. Because existing works commonly use a split proposed by Eigen et al. [11] for the training and testing, we also follow it to compare with those works. Therefore, 697 images covering a total of 29 scenes are used for evaluation, and the remaining 32 scenes of 23,488 images are used for the training. We use $\kappa = 80$ for this dataset.

4.4. Evaluation Result

For evaluation, we use following metrics used by previous works:

Threshold : % of \tilde{d}_i s.t. $\max(\frac{\tilde{d}_i}{d_i}, \frac{d_i}{\tilde{d}_i}) = \delta < thr$,

Abs Rel : $\frac{1}{|T|} \sum_{\tilde{d} \in T} |\tilde{d} - d|/d$,

Sq Rel : $\frac{1}{|T|} \sum_{\tilde{d} \in T} \|\tilde{d} - d\|^2/d$, RMSE :

$\sqrt{\frac{1}{|T|} \sum_{\tilde{d} \in T} \|\tilde{d} - d\|^2}$,

log10 : $\frac{1}{|T|} \sum_{\tilde{d} \in T} |\log_{10} \tilde{d} - \log_{10} d|$, RMSElog :

$\sqrt{\frac{1}{|T|} \sum_{\tilde{d} \in T} \|\log \tilde{d} - \log d\|^2}$, where T denotes a collection of pixels that the ground truth values are available.

Using NYU Depth V2 dataset, the experimental results given in Table 1 show that Ours-DenseNet achieves the state-of-the-art result with a significant margin in both of the inlier measures (*i.e.*, $\delta < thr$) and accuracy metrics (*i.e.*, AbsRel, log10) except only RMSE. Our ResNet-based model also outperforms the method from Yin et al [51] with a significant margin which has the same backbone network,

ResNext-101 [48].

In the evaluation using the KITTI dataset provided in Table 2, ours outperforms all existing works with a significant margin. Please note that our ResNet-based model already outperforms the methods from Fu et al [12] and Yin et al [51] which use ResNet-101 and ResNext-101 [48] as their backbone network, respectively. Only the root mean squared errors (RMSE) from ours are behind that of Fu et al.’s in the capturing range 0-80m. However, in the capturing range 0-50m, ours-ResNet achieves more than 10% improvement in RMSE from the result of Fu et al. Also, the proposed method achieves notable improvements in the inlier metrics (*i.e.*, $\delta < thr$), meaning more number of correctly estimated pixels as it can be seen from Figures 5 and 6 presenting qualitative comparison to our competitors.

We also evaluate the proposed method on the online KITTI benchmark server with a model trained using KITTI’s official split. Apart from the training set, all other settings remain the same as in the experiment using KITTI’s Eigen split. We trained Ours-DenseNet for 50 epochs with 28,654 image-ground truth pairs sampled from the official training and validation set. As shown in Table 3, our method outperforms all the published works.

4.5. Ablation Study

Here, we conduct evaluations with variants of our network to see the effectiveness of the core factors. From the baseline network, which only consists of the base network (*i.e.*, *ResNet-101*) and a simple upsampling layer, we increment the network with the core modules to see how the added factor improves the performance. The result is given in Table 4. As the core factors are added, the overall performance is improved, and the most significant improvement is made by adding the proposed local planar guidance layers. Please note that the LPG layers only require additional 0.1M trainable parameters used by *1x1 reduction* layers. The final improvement comes from using the training loss defined in Equation 5. Benefited from the robust base

Variant	# Params	<i>higher is better</i>			<i>lower is better</i>				
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE <i>log</i>	log10
MobileNetV2 [37]	16.3M	0.860	0.974	0.993	0.121	0.080	0.431	0.156	0.052
ResNet-50 [18]	49.5M	0.865	0.975	0.993	0.119	0.075	0.419	0.152	0.051
ResNet-101 [18]	68.5M	0.871	0.977	0.995	0.113	0.068	0.407	0.148	0.049
ResNext-50 [48]	49.0M	0.867	0.977	0.995	0.116	0.070	0.414	0.150	0.050
ResNext-101 [48]	112.8M	0.880	0.977	0.994	0.111	0.069	0.399	0.145	0.048
DenseNet-121 [20]	21.2M	0.871	0.977	0.993	0.118	0.072	0.410	0.149	0.050
DenseNet-161 [20]	47.0M	0.885	0.978	0.994	0.110	0.066	0.392	0.142	0.047

Table 5: Experimental results using NYU Depth V2 with various base networks.

Variant	# Params	<i>higher is better</i>			<i>lower is better</i>				
		$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Sq Rel	RMSE	RMSE <i>log</i>	log10
ResNet-50 [18]	49.5M	0.954	0.992	0.998	0.061	0.250	2.803	0.098	0.027
ResNet-101 [18]	68.5M	0.954	0.993	0.998	0.061	0.261	2.834	0.099	0.027
DenseNet-121 [20]	21.2M	0.951	0.993	0.998	0.063	0.256	2.850	0.100	0.028
DenseNet-161 [20]	47.0M	0.955	0.993	0.998	0.060	0.249	2.798	0.096	0.027
ResNext-50 [48]	49.0M	0.954	0.993	0.998	0.061	0.245	2.774	0.098	0.027
ResNext-101 [48]	112.8M	0.956	0.993	0.998	0.059	0.241	2.756	0.096	0.026

Table 6: Experimental results using KITTI’s Eigen split with various base networks. In this experiment, we set the capturing range to 0 – 80m.

network *DenseNet-161*, ours achieves state-of-the-art performance with the significant margin while it requires the less number of parameters than the baseline.

4.6. Experiments with Various Base Networks

Because the proposed network adopts existing models as an encoder for dense feature extraction, it is desirable to see how the performance varies with various base networks that are widely used for similar applications. By changing the encoder with various models while other settings remained, we experimented with the proposed method using both of the NYU Depth V2 and KITTI’s Eigen split, and provide the result in Tables 5 and 6. Note that ResNet-101, ResNext-101 and DenseNet-161 are identical to Ours-ResNet, Ours-ResNext and Ours-DenseNet, respectively, in Tables 1 and 2. Interestingly, for the NYU Depth V2 dataset, DenseNet-161 results in the best performance, while for KITTI’s Eigen split, ResNext-101 achieves the state-of-the-art result. We consider this as an effect of the relatively lower variance of the data distribution in the indoor scenes of the NYU Depth V2 dataset, which can lead to a degeneration of performance with very deep models such as ResNext-101 in this experiment. Also, it is notable that our MobileNetV2-based model results in performance drop about only 3% for inlier measures and less than 15% drop for accuracy measures while it contains less than half the number of parameters and shows about three times speedup when compared to our model based on the DenseNet-161.

4.7. Qualitative Result

Finally, we discuss qualitative results from ours and competing works. As we can see from Figures 5 and 6, ours show much more precise object boundaries. However, in results from experiments using KITTI, we can see artifacts in the sky or upper part of the scenes. We consider this as a consequence of the very sparse ground truth depth data. Because certain regions are lacking valid depth values across the dataset, the network cannot be appropriately trained for those regions.

5. Conclusion

In this work, we have presented a supervised monocular depth estimation network and achieved state-of-the-art results. Benefiting from recent advances in deep learning, we design a network architecture that uses novel local planar guidance layers, giving an explicit relation from internal feature maps to the desired prediction for better training of the network. By deploying the proposed layers on multiple stages in the decoding phase, we have gained a significant improvement and shown several experimental results on challenging benchmarks to verify it. However, in experiments with the KITTI dataset, we observe frequent artifacts on the upper part of the scenes. We analyze this as an effect of the high sparsity of the ground truth across the dataset. As a consequence, we plan to investigate adopting into our framework a photometric reconstruction loss, which can provide far denser supervision to improve the performance further.

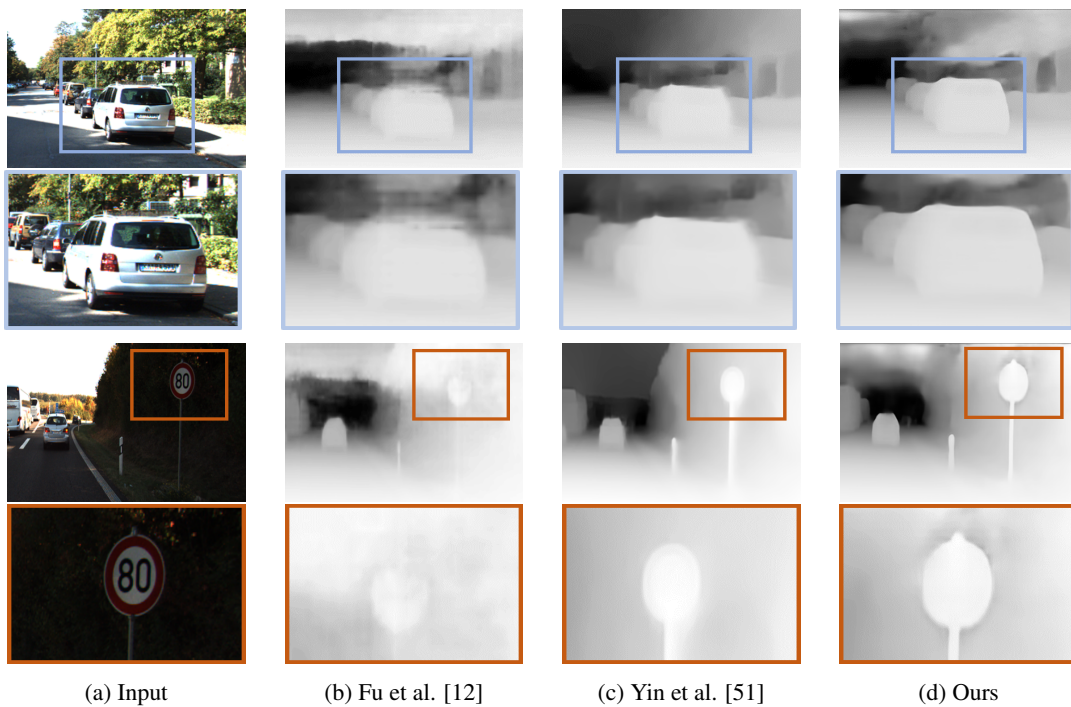


Figure 5: **Qualitative results on the KITTI Eigen test split.** The proposed method results clearer boundaries from the vehicle and traffic sign.

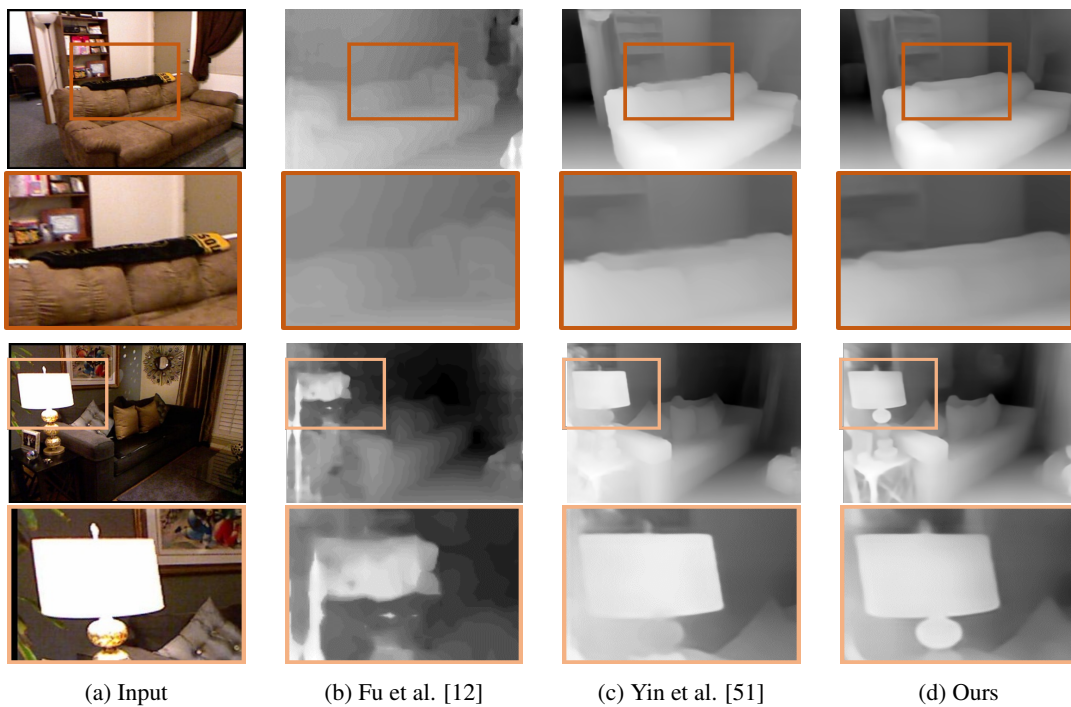


Figure 6: **Qualitative results on the NYU Depth V2 test split.** While the method from Yin et al [51] show competitive results to ours, our method achieves more distinctive results especially on object boundaries.

References

- [1] A. Abrams, C. Hawley, and R. Pless. Heliometric stereo: Shape from sun position. In *Computer Vision—ECCV 2012*, pages 357–370. Springer, 2012.
- [2] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007.
- [3] Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
- [4] A. Chakrabarti, J. Shao, and G. Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. In *Advances in Neural Information Processing Systems*, pages 2658–2666, 2016.
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.
- [6] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*, pages 730–738, 2016.
- [7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
- [8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [9] R. Diaz and A. Marathe. Soft labels for ordinal regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [10] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
- [11] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.
- [12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.
- [13] Y. Gan, X. Xu, W. Sun, and L. Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 224–239, 2018.
- [14] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.
- [15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [16] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.
- [17] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [19] I. P. Howard. *Perceiving in depth, volume 1: basic mechanisms*. Oxford University Press, 2012.
- [20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.
- [21] S. Kim, K. Park, K. Sohn, and S. Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *European conference on computer vision*, pages 143–159. Springer, 2016.
- [22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [23] P. Knöbelreiter, C. Reinbacher, A. Shekhovtsov, and T. Pock. End-to-end training of hybrid cnn-crf models for stereo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [24] Y. Kuznetsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017.
- [25] 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*, pages 239–248. IEEE, 2016.
- [26] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 330–339, 2018.
- [27] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.
- [28] J. Li, R. Klein, and A. Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017.
- [29] F. Liu, C. Shen, G. Lin, and I. D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, 2016.
- [30] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.
- [31] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

- [32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshine, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [33] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.
- [34] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4058–4066, 2016.
- [35] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.
- [36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [38] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.
- [39] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009.
- [40] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.
- [41] A. G. Schwing and R. Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.
- [42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.
- [43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [44] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.
- [45] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.
- [46] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.
- [47] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016.
- [48] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [49] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.
- [50] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.
- [51] W. Yin, Y. Liu, C. Shen, and Y. Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [52] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.