Journal of Real-Time Image Processing (2021) 18:1357–1368
https://doi.org/10.1007/s11554-021-01092-0

SPECIAL ISSUE PAPER

Efficient unsupervised monocular depth estimation using attention guided generative adversarial network

Sumanta Bhattacharyya[1] · Ju Shen[2] · Stephen Welch[1] · Chen Chen[1]

Received: 31 January 2020 / Accepted: 6 March 2021 / Published online: 22 March 2021

Abstract
Deep-learning-based approaches to depth estimation are rapidly advancing, offering better performance over traditional computer vision approaches across many domains. However, for many critical applications, cutting-edge deep-learning based approaches require too much computational overhead to be operationally feasible. This is especially true for depth-estimation methods that leverage adversarial learning, such as Generative Adversarial Networks (GANs). In this paper, we propose a computationally efficient GAN for unsupervised monocular depth estimation using factorized convolutions and an attention mechanism. Specifically, we leverage the Extremely Efficient Spatial Pyramid of Depth-wise Dilated Separable Convolutions (EESP) module of ESPNetv2 inside the network, leading to a total reduction of 22.8%, 35.37%, and 31.5% in the number of model parameters, FLOPs, and inference time respectively, as compared to the previous unsupervised GAN approach. Finally, we propose a context-aware attention architecture to generate detail-oriented depth images. We demonstrate superior performance of our proposed model on two benchmark datasets KITTI and Cityscapes. We have also provided more qualitative examples (Fig. 8) at the end of this paper.

Keywords Attention · Efficient GAN · Unsupervised depth estimation · Convolution factorization

1 Introduction

Image-based depth estimation is a key problem in computer vision, with a wide range of applications from robotic navigation to virtual reality. Early applications of deep learning to depth estimation used to rely on supervised learning, directly regressing a depth estimate of each pixel, and training models using ground-truth depth maps. Eigen et al. [1] has demonstrated good performance using a multi-scale Convolutional Neural Network (CNN) to predict depth from single images. In order to learn the pixel-wise transformation, supervised approaches require ground truth depth data for training. However, obtaining the ground truth depth data is non-trivial, and model performance may be limited by the amount of quality ground truth data that can be collected. Probabilistic graphical models such as Conditional Random Field [2] have increased the performance when they are used in neural networks for optimization.

Unsupervised approaches estimate disparity maps from two different image views (rectified left and right images) of the calibrated stereo camera, making ground truth depth data not required for training. This makes the unsupervised approaches more robust in practice. Godard et al. [3] proposed a left-right cycle consistency loss as a constraint on this unsupervised approach. Pilzer et al. [4] apply the left-right consistency in an adversarial learning approach in order to improve the generated images. Although adversarial learning-based unsupervised methods achieve excellent performance in depth estimation, these methods rely on the complex generative adversarial network (GAN) architecture which is generally computationally heavy. As a result, they are not able to run in real-time on resource-constrained edge devices for practical applications, e.g. autonomous driving.

✉ Sumanta Bhattacharyya
sbhatta9@uncc.edu

Ju Shen
jshen1@udayton.edu

Stephen Welch
stephencwelch@gmail.com

Chen Chen
chen.chen@uncc.edu

[1] University of North Carolina at Charlotte, Charlotte, USA

[2] University of Dayton, 300 College Park, Dayton, OH 45469, USA

To address this challenge, we propose a computationally efficient architecture for depth estimation given stereo image pairs, based on the unsupervised GAN architecture [4]. A context-aware attention mechanism is also introduced to improve depth estimation, yielding more accurate overall depth prediction. In summary, the main contributions of this paper are:

1. We adopt the EESP module [5] inside a GAN architecture to significantly improve the computational efficiency, while concurrently reducing RMSE by 7% compared to the baseline model [4].
2. We introduce a context aware attention layer in the generator to get accurate depth images. To our knowledge, our work is the first to explore the attention mechanism in the unsupervised depth estimation approach.
3. We conduct extensive experiments on the publicly available datasets KITTI and CityScapes. The results reveal the effectiveness of the proposed method. A detailed ablation study is also carried out to identify the relative contributions of the individual components.

2 Related work

Image-based depth estimation techniques estimate the 3D structure of a scene from a 2D image. In traditional computer vision, depth estimation algorithms rely on point correspondences between stereo image pairs [6, 7] and triangulation. Saxena et al. [8] show that depth can be learned from handcrafted features using monocular cues of a single 2D image. These approaches have evolved over time [8–11], and are in many cases today being displaced by deep learning based approaches, which no longer require the handcrafted features, but instead estimate depth directly from raw pixel values. Here we focus on models that take in a single input image and predict the depth of the image.

Supervised depth estimation Supervised learning relies on ground truth depth data to achieve promising performance for image depth estimation. Indoor datasets like NYU [12] and outdoor datasets like KITTI [13] and Cityscapes [14] contribute to the evolution of supervised monocular depth estimation approaches. Eigen et al. [1] propose a two-scale network to generate a dense depth map trained on ground truth values. Probabilistic graphical models (i.e. MRF and CRF) have also been combined with deep networks to boost accuracy [15]. Xu et al. [2] offered structured attention mechanism using CRF to combine multi-scale information obtained from the CNN layers. Although supervised depth estimation has traditionally been formulated as a regression problem, Cao et al. [16] show that there may be advantages to formulate the task as a pixel-wise classification problem. Recent architectures have shown promising developments for multi-task learning strategies [17, 18] that include depth estimation. Chen et al. [19] utilize two different depth estimation methods (coarse-level depth estimation and saliency-aware depth enhancement) in order to get better saliency detection. This saliency-aware depth enhancement generates different depth values for salient and non-salient objects. These approaches are often quite complex and rely on ground truth depth for training. In contrast, our approach does not require ground truth depth value while training.

Unsupervised depth estimation Recent unsupervised depth estimation algorithms [20, 21] have gained popularity in the research community. Garg et al. [22] introduce an unsupervised approach using stereo image pairs. However, the method is limited by a loss function that is not fully differentiable. Albeit a linearization of the loss function via Taylor approximation is developed which is still challenging to optimize. Zhou et al. [23] solve this problem by using bilinear warping. In a recent work, Godard et al. [3] propose a new left right cycle consistency loss along with the image reconstruction loss for a higher quality depth estimation. This new training loss for depth estimation has also been used in Cycle-GAN architecture by Pilzer et al. [4] for unsupervised depth estimation based on stereo disparity estimation. Although these approaches have achieved impressive results in unsupervised manner, it is difficult for GAN [24] architectures to achieve real-time throughput on resource constrained embedded devices due to their high complexity. Our work attempts to alleviate this problem by introducing an efficient GAN framework for unsupervised depth estimation.

Adversarial learning Adversarial learning has been proven to be efficient in the image generation task. A line of research has explored various GAN architectures [24] for dense depth map generation. For example, Kundu et al. [25] leverage the domain adaptation strategy for depth estimation in an adversarial learning framework. Various GAN architectures such as Conditional GAN [26], CycleGAN [27] are utilized for depth estimation task [11]. Joint learning strategy [28, 29] has also been investigated through GANs for high quality depth estimation. In this paper, we focus on building cost effective GAN architecture, which is significantly different than the previous works.

Attention Attention models are very useful in computer vision for improving the performance in pixel-level prediction tasks as well as in the context of monocular depth prediction. Depth maps need to be accurate and detail oriented, so preserving details through attention layers may prove helpful in 3D reconstruction. There are several applications of attention layers to supervised depth estimation. Hao et al. [30] demonstrate how attention mechanism can focus on the most informative part of the input image based on the context. Chen et al. [11] also use attention as aggregation of image and pixel level information. The approach we present

in this paper is different from the prior works in the following two aspects. First, ours is the first work to explore the attention mechanism in the unsupervised depth estimation problem. Second, we leverage the advantage of multi-scale feature fusion (local and global) to obtain attention aware features, leading to enhanced depth estimation.

3 Method

In this section we first describe the baseline architecture along with our proposed method and then explain each component of this proposed architecture. In the later sections, we demonstrate its efficiency in image generation.

3.1 Network architecture

We follow Pilzer et al.'s [4] work on Cycle-GAN architecture for depth estimation as the baseline of our approach. As shown in Fig. 1, this architecture uses calibrated stereo camera images (pairwise) as input to estimate disparity map (d_m) through image synthesis. The generator network consists of two sub-networks. The upper sub-network generates a right disparity map (R_d) with the input I_l and synthesizes a right image view (I'_r) through the warping operation (change pixels locations to create a new image) $\mathbf{w}$, $I'_r = \mathbf{w}(R_d, I_l)$. Similarly, the lower sub-network generates a left image view, $I'_l = \mathbf{w}(L_d, I_r)$. The reconstruction loss (L_r) is implemented between the synthesized and input images in order to optimize the generator networks:

$$L_r = \left\|I_r - \mathbf{w}(R_d, I_l)\right\| + \left\|I_l - \mathbf{w}(L_d, I'_r)\right\|. \tag{1}$$

The discriminator, $D1$, $D2$, is used to discriminate if the synthesized image, I'_l, I'_r, is fake or not, thus the adversarial loss can be formulated as

$$\begin{aligned} L_{\text{GAN}} = {} & \mathbb{E}_{I_r \sim P(I_r)}[\log D1(I_r)] \\ & + \mathbb{E}_{I_l \sim P(I_l)}[\log(1 - D1(\mathbf{w}(R_d, I_l)))] \\ & + \mathbb{E}_{I_l \sim P(I_l)}[\log D2(I_l)] \\ & + \mathbb{E}_{I_r \sim P(I_r)}[\log(1 - D2(\mathbf{w}(R_d, L_d)))]. \end{aligned} \tag{2}$$

Each half generates disparities of different views, R_d, L_d. To enforce a view constraint, a consistency loss is formulated,

$$L_c = \left\|L_d - \mathbf{w}(L_d, R_d)\right\|. \tag{3}$$

We consider structural similarity loss (L_{ssim}) along with the adversarial loss for better full-cycle optimization.

$$L_{\text{ssim}} = \frac{(2 \times \mu_x \times \mu_y + C_1) \times (2 \times \sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4}$$

where μ_x and μ_y denote local sample means of x and y respectively. σ_x and σ_y denote local sample standard deviations of x and y respectively. σ_{xy} denotes local sample correlation coefficient between x and y. C_1 and C_2 are constants

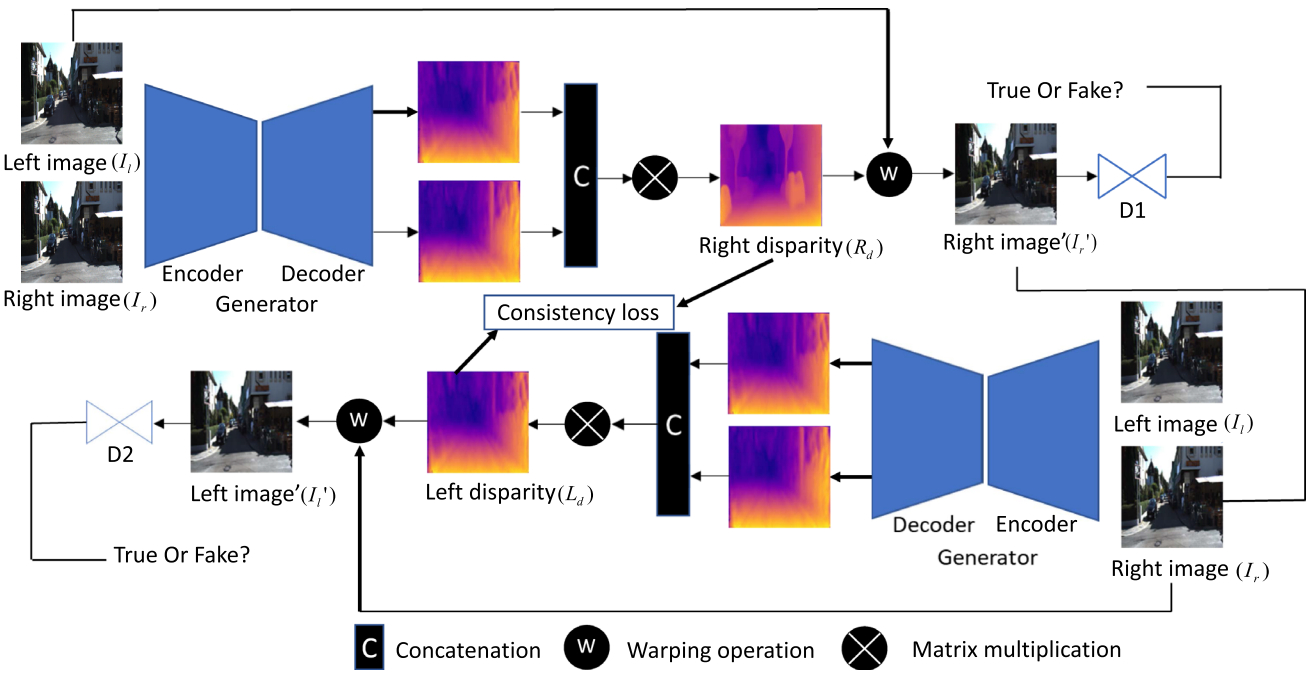

Fig. 1 Unsupervised monocular depth estimation framework using Cycle-GAN

for stabilization if denominator is too small. The total loss for the full cycle optimization is:

$$\textbf{Loss} = \alpha L_r + \beta L_{\text{GAN}} + \gamma L_c + \delta L_{\text{ssim}}. \quad (5)$$

α, β, γ and δ are the corresponding weights for different losses. The output disparity map is obtained by

$$\mathbf{D} = (L_d + \mathbf{w}(L_d, R_d))/2. \quad (6)$$

As shown in Fig. 2, the generator and discriminator parts in the original architecture have been modified by replacing the standard convolutions with EESP convolution factorization and implemented the proposed context aware attention layer in the decoder part of generator network. As proposed by Zhang et al. [31], attention mechanism works best on middle to high level feature maps as it receives more evidence to choose the conditions. Attention layer has been applied to the first two layers of decoder since they contain the high level feature maps of the generated image. In order to stabilize learning, spectral normalization [32] in the discriminator network has been used, as it is critical for the generator to learn the multi-modal structure of the target distribution by controlling the performance of a discriminator. Spectral normalization puts constrains on the Lipschitz constant of the discriminator network without any extra hyper-parameter tuning [32] to improve GAN training stability.

3.2 Explanation of different convolution methods

As mentioned in Table 1, we compare with different type of convolutions in order to demonstrate the efficiency of EESP module along with the original motivation of selecting specific type of convolution methods.

A standard convolution is the element-wise multiplication and addition. Dilated convolution uses dilation rate. It defines a spacing between the values in a kernel (i.e. 3×3 kernel with a dilation rate of 2 will have the same receptive field size as a 5×5 kernel, while preserving the same number of parameters). Depthwise convolution applies a single convolutional filter for each input channel and uses pointwise convolution to create a linear combination of the output of the depthwise convolution. Depthwise Dilated Separable Convolution includes both depthwise separability and dilated convolution. Grouped convolution applies a group of convolutions. It uses multiple kernels per layer which results in multiple channel outputs per layer. This leads to good learning of low-level and high-level features. AlexNet [33] applied Grouped convolution for model distribution over multiple GPUs (Fig. 4).

3.3 EESP module

The EESP module is empowered by group convolution and parallel branches of depth-wise dilated separable convolution. As shown in Fig. 5a, this technique reduces the high dimensional input features into a low dimensional space using group convolution. Then it learns the low dimensional feature representation in parallel branches using depth-wise

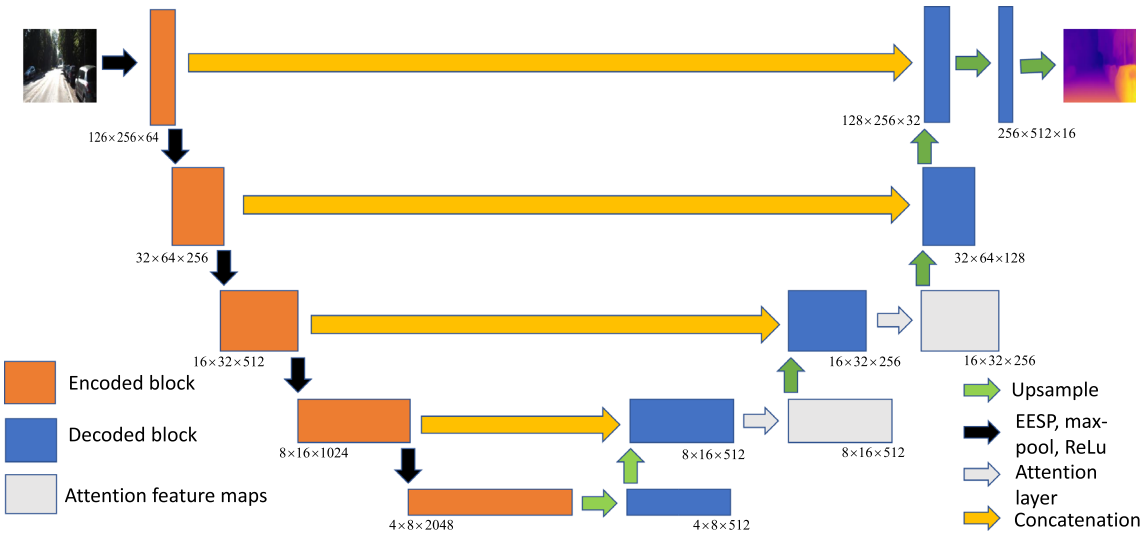

Fig. 2 Decomposition of the generator part in Fig. 1. Our method implements context-aware attention block in the decoder part of generator along with EESP unit for efficiency and accuracy. We apply the attention mechanism in the early stages of the decoder to better extract medium to high level features. The attention mechanism described by grey blocks in this figure has been explained in detail in Fig. 3

dilated separable convolution with different dilation rates (larger dilation rate corresponds to larger size of receptive field) followed by hierarchical addition in order to remove the gridding artifacts. As proposed by Mehta et al. [5] depthwise dilated separable convolutions and group convolutions are more efficient.

3.4 Attention layer

Convolution layers process images in a local neighborhood of the image. Convolution layers alone do not capture long range dependencies. These long range dependencies are useful in generative networks like GAN to enhance the synthesized images. In this section, we will discuss about our context aware lightweight attention mechanism, as shown in

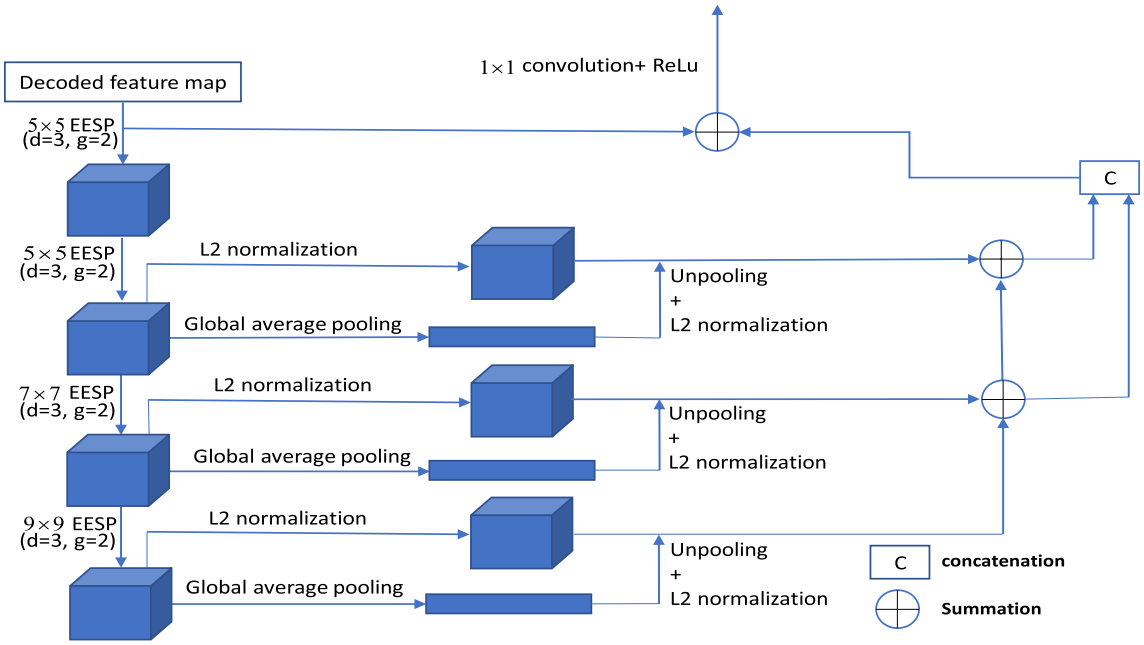

Fig. 3 The context-aware attention architecture. Increasing receptive field of kernels [(with dilation rate (d) = 3 and group (g) = 2] and their corresponding global feature context helps to obtain context aware attention features. The dilation rate and group number are fixed in the architecture for the factorized convolution. Intermediate decoded feature map is the input to the architecture as shown in Fig. 2. In our case, hierarchical fusion of different layer features provides the best result

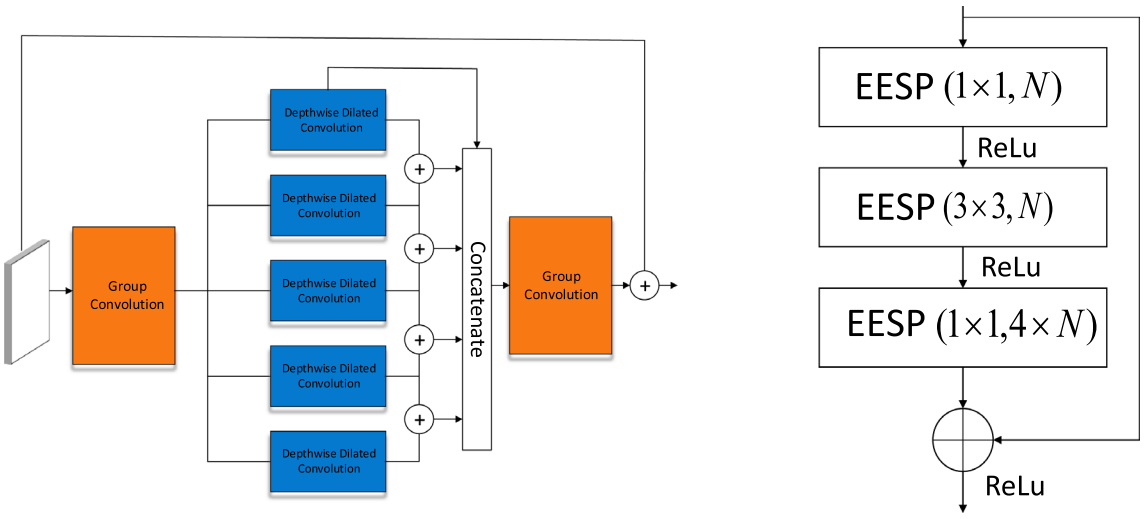

Fig. 4 Convolution factorization block. **a** Schematic diagram of a single EESP unit. **b** A bottleneck building block [34] using EESP (N = output dimension)

Table 1 Comparison of different convolution operations with a 3×3 kernel, M input channels, and N output channels

Convolution type	Parameter #	Receptive field size
Standard convolution	$3 \times 3 \times M \times N$	3×3
Group convolution	$\frac{3 \times 3 \times M \times N}{groups}$	3×3
Dilated convolution	$3 \times 3 \times M \times N$	$d_k \times d_k$
Depth-wise dilated separable convolution	$3^2 \times M + M \times N$	$d_k \times d_k$

For dilated convolution, 3×3 kernel is used with a dilation rate d. The receptive field size is computed as $d_k = ((3-1) \times d + 1)$

Fig. 3. This attention module is able to capture the detailed context information to enhance the estimated depth image.

Our attention mechanism exploits multi-scale features fusion [35] for each layer with increasing receptive field of kernels. We also use factorized convolution (EESP module) for feature extraction. These features are concatenated with their corresponding global context features. Specifically, the global context features are obtained through global average pooling (channel wise attention) across the whole feature map for each layer. Finally, the multi-scale outputs are hierarchically fused to generate the final output feature map.

Global average pooling outputs a 1D context vector which is replicated to the same size of the feature maps to merge. Merging two features for each scale is not efficient enough to produce a good result because (1) different scales of the two feature maps and (2) the unpooled global feature vector is not as dominant as large multi-scale feature maps, so plain concatenation may be futile. Although during training the weight might get adjusted, it requires heavy parameter tuning. So we apply per-pixel L_2 normalization to both features to be merged along with a learnable scale parameter for each channel. For an n-dimensional input X after L_2 normalization we obtain $X_1 = \zeta * \frac{X}{\|\mathbf{X}\|_2}$, where $\|\mathbf{X}\|_2$ is $\sqrt{\sum_{n=1}^{d} |X_n|^2}$ and ζ is a scaling parameter.

Table 2 Efficiency comparison between two architectures

Network	Arch.	FLOPs (bil.)	Param (mil.)
Original GAN [4]	GAN	8993	125
Original GAN+EESP [4]	GAN	8343	91.2
Original GAN+attention [4]	GAN	9213	129.5
Ours	GAN	**5833**	**96.5**

The best scores are marked in bold

Our approach achieves a significant reduction of computational complexity as compared to the original GAN approach [4]

Table 3 Total inference time using a single NVIDIA-GTX 1080Ti GPU on the KITTI dataset

Network	Architecture	Inf. time (s)
Original GAN [4]	GAN	0.190
Original GAN+EESP [4]	GAN	0.115
Original GAN+attention [4]	GAN	0.228
Ours	GAN	**0.130**

The best scores are marked in bold

It shows our approach outperforms the baseline model [4]

4 Experiments

In this section, we evaluated our proposed method extensively using KITTI [13] and Cityscapes [14] datasets. We present quantitative and qualitative results to demonstrate the effectiveness of the proposed model (Tables 2 and 3).

4.1 Dataset

KITTI dataset [13] contains several outdoor scenes from LIDAR sensor and car-mounted cameras while driving. We use the data split as suggested by Eigen et al. [1] for both training and testing. It contains 22600 training image pairs and 697 test image pairs. The input images have been down sampled to 512×256 resolution image with respect to original resolution of 1224×368. Random data augmentation has been done by flipping of images during training. Cityscapes dataset [14] consists of 22,973 stereo image pairs for training captured across various German cities. It gives higher resolution image quality and variety compared to KITTI. Both of these datasets are highly recognized for various computer vision tasks, segmentation, classification, depth prediction *etc.*

4.2 Implementation details

In our experiments, we set the dilation rate d in EESP module proportional to the number of branches in the EESP (for our experiments, we used 5 parallel branches, dilation rates from 2^0 to 2^4, with number of groups 2). The effective receptive field of the EESP unit grows with the number of branches, as shown in Fig. 4a. As shown in Fig. 2, the generator networks use Resnet-50 network for the encoder and the decoder contains five deconvolutional layers with ReLu activation function. For the first two layers in the decoder, we integrate the attention layer in order to process the large feature maps for context information. Skip connections are used to pass information from encoder to decoder in order to aggregate efficient feature representation. All the convolution operations in the generator part are replaced by the

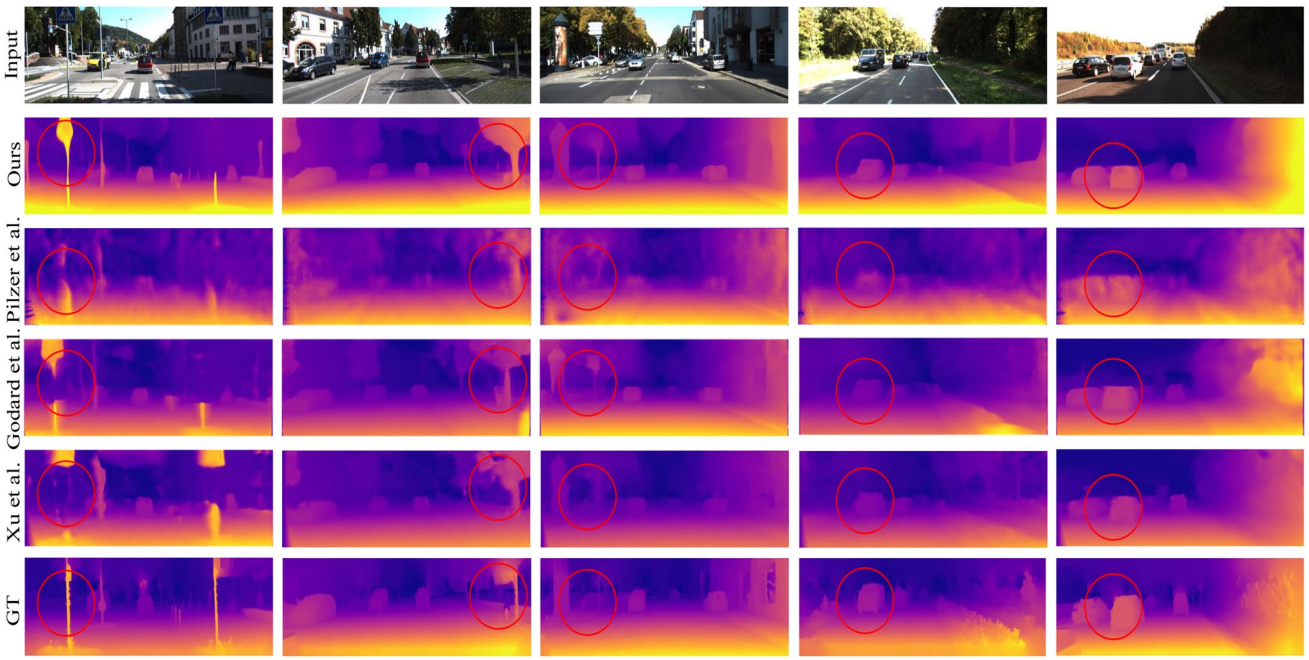

Fig. 5 Qualitative measurements on KITTI dataset [13]. Due to the attention layer, our approach generates the subtle structural details of the image compared to the other state of the art unsupervised methods. The ground truth depth maps are interpolated from sparse LIDAR points for visualization purpose only

factorized EESP module. $D1$ and $D2$ each has five consecutive EESP operations. We use bilinear sampler for the warping operation.

4.3 Experimental setup

The proposed method is implemented using TensorFlow [36] and takes 21 hours to train using a single NVIDIA-GTX 1080Ti GPU. The batch size is set to 8. The initial learning rate is 10^{-6} and is reduced by half at [40k, 70k] steps. We use ADAM [37] optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.9$ and weight decay = 0.006 to train the model with 100 epochs.

4.4 Evaluation

We evaluate our depth estimation using the following evaluation metrics. Considering d_i and d_{gi} are the estimated depth and ground truth depth value for pixel i. T is the total number of valid pixels in the test set.

$$\text{Abs. Rel} = \frac{1}{T}\sum_i \frac{d_i - d_{gi}}{d_{gi}} \tag{7}$$

$$\text{Sq. Rel} = \frac{1}{T}\sum_i \frac{|d_i - d_{gi}|^2}{d_{gi}} \tag{8}$$

$$\text{log RMSE} = \sqrt{\frac{1}{T}\sum_i \left\|(\log(d_i) - \log(d_{gi}))\right\|^2} \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_i \|(d_i - d_{gi})\|^2} \tag{10}$$

The accuracy with threshold t so that $\delta = \max(\frac{d_{gi}}{d_i}, \frac{d_i}{d_{gi}}) < t$, where t=1.25, 1.25^2, 1.25^3.

We compare our proposed model with the state of the art supervised and unsupervised depth estimation methods for both datasets. Tables 4 and 5 and Figs. 5 and 6 are the respective quantitative and qualitative analysis of our method and other approaches on KITTI and Cityscapes. In comparison with the supervised approaches, we have achieved very similar results to the best performing method i.e. Xu et al. [39]. In case of unsupervised approaches, our approach significantly outperforms Godard et al. [3], which represents the state of the art among unsupervised approaches to this task. Finally, we also compare with Pilzer et al.'s [4] full-cycle+D training and ours yield better results.

4.5 Ablation study

To validate the contribution of our context aware attention strategy and the convolution factorization to overall performance, we present an ablation study on KITTI dataset, i.e.

Table 4 Quantitative Comparison with the state of the art methods trained and tested on KITTI dataset. Supervised and unsupervised methods are labelled as "Y" and "N"

Method	Sup	Abs.Rel ↓	Sq.Rel ↓	RMSE ↓	RMSE(log) ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Eigen et al. [1]	Y	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Liu et al. [38]	Y	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Xu et al. [39]	Y	*0.132*	*0.911*	–	**0**.162	0.804	**0**.945	*0981*
Zhou et al. [40]	N	0.208	1.768	6.856	0.283	0.678	0.885	0.957
AdaDepth et al. [25]	N	0.203	1.734	6.251	0.284	0.687	0.899	0.958
Garg et al. [22]	N	0.169	1.080	*5.104*	0.273	0.740	0.904	0.962
Godard et al. [3]	N	0.148	1.344	5.927	0.247	*0803*	0.922	0.964
Pilzer et al. [4]	N	0.198	1.990	6.655	0.292	0.721	0.884	0.949
Wang et al. [21]	N	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Ours (EESP+attention)	N	**0**.1196	**0**.889	**4**.329	*0.192*	**0**.865	*0.943*	**0**.989

Results are obtained on Eigen split dataset. We train the Pilzer et al. [4] method using the Full-Cycle+D method. Bold indicates the best result and italics is the second best. Our approach outperforms the state of the art on 5 out of 7 evaluation metrics

Table 5 Quantitative Comparison on Cityscapes dataset. Supervised and unsupervised methods are labeled as "Y" and "N"

Method	Sup	Abs.Rel ↓	Sq.Rel ↓	RMSE ↓	RMSE(log) ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Pilzer et al. [4]	N	0.440	6.036	5.443	0.398	0.730	0.887	0.944
Wang et al. [21]	N	0.148	1.187	5.496	0.226	0.812	0.938	0.975
Godard et al. [3]	N	*0.097*	*0.896*	*5.093*	**0**.176	**0**.879	*0962*	**0**.986
Zhou et al. [40]	N	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Ours (EESP+attention)	N	**0**.090	**0**.813	**4**.633	*0.193*	*0.832*	**0**.974	*0.978*

We train Pilzer et al. [4]'s network using the Full-Cycle+D method. Bold indicates the best result and italics is the second best. Our approach outperforms existing state of the art approaches on 4 out of 7 evaluation metrics. Note that we directly apply our model trained on KITTI dataset without any specific tuning

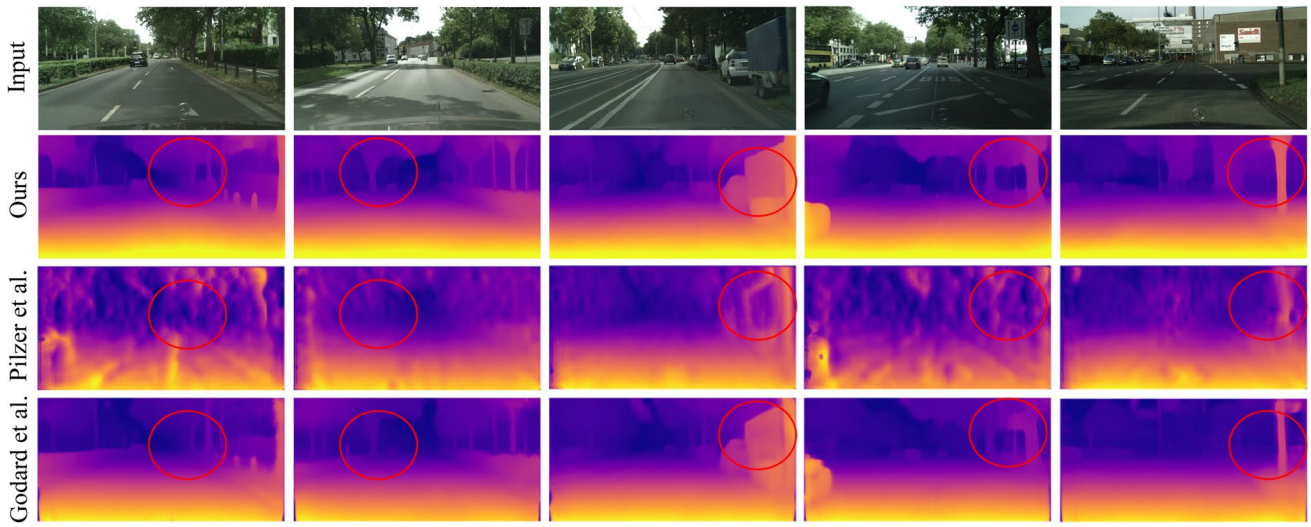

Fig. 6 Qualitative comparison on the Cityscapes dataset. Our model is trained on KITTI dataset and evaluated on Cityscapes without any specific tuning

Table 6 The components in our proposed model for ablation study

Methods	Convolution operation	EESP operation	Context aware attention mechanism
Baseline [4]	✓	×	×
Ours (EESP)	×	✓	×
Ours (attention)	✓	×	✓
Ours (EESP+attention)	×	✓	✓

(1) replacing convolution operations with EESP (2) implementation of attention mechanism to the baseline with standard convolution operations. Table 6 shows the breakdown of various components involved in each experiment. This highlights the impact of each component (i.e. EESP, attention layer) on the baseline model.

As can be seen from Table 7, the model with only EESP operation performs a little better than the baseline model. However, the factorization operation leads to significant ease of computation. Then, we implement our attention mechanism with baseline which clearly generates an enhanced depth map. This verifies our intuition of integrating attention with EESP operation to improve the generated depth image.

As illustrated in Fig. 7, we qualitatively demonstrate the impact of context-aware attention model along with the impact of EESP operation alone by presenting the predicted depth maps from initial training epochs. The evolution of these predicted depth maps reveal that our context aware attention architecture is able to focus on the salient objects in the image and captures the depth information at the early stage (e.g. the first epoch) of training compared to both the baseline architecture and EESP+baseline architecture. It shows that attention mechanism improves the architecture's ability to learn finer image details.

Table 7 Quantitative evaluation of different modifications of the network on KITTI dataset as ablation study

Method	Sup	Abs.Rel ↓	Sq.Rel ↓	RMSE ↓	RMSE(log) ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Baseline [4]	N	0.198	1.990	6.655	0.292	0.721	0.884	0.949
Ours (EESP)	N	0.195	1.76	6.09	0.292	0.758	0.905	0.958
Ours (attention)	N	*0.138*	*0.915*	*4.571*	*0.247*	*0.831*	*0.919*	*0.964*
Ours (EESP+attention)	N	**0.1196**	**0.889**	**4.329**	**0.192**	**0.865**	**0.943**	**0.989**

It is evident that the attention mechanism is able to improve the performance as compared to using EESP module alone. Bold indicates the best result and italics is the second best

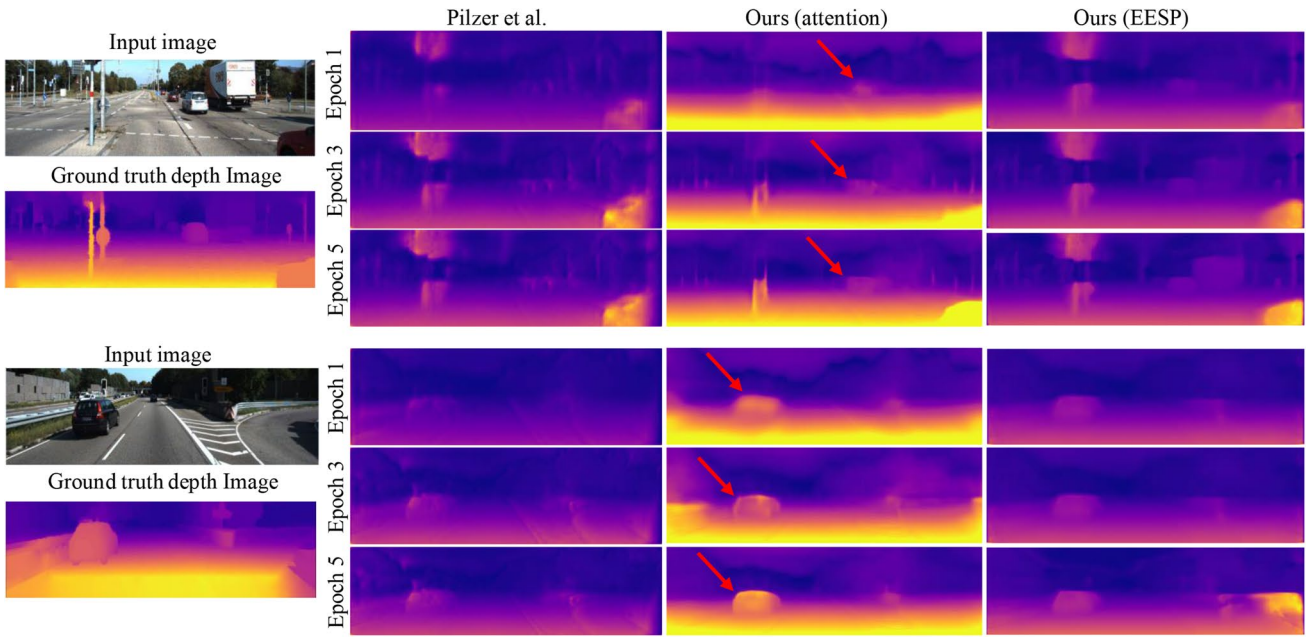

Fig. 7 Qualitative comparison of the ablation study. Our proposed context-aware attention mechanism provides effective learning and captures refined image details in the early stage of learning. For example, it learns the structure of the far-away car in epoch 1 for both images while other two networks failed to do that

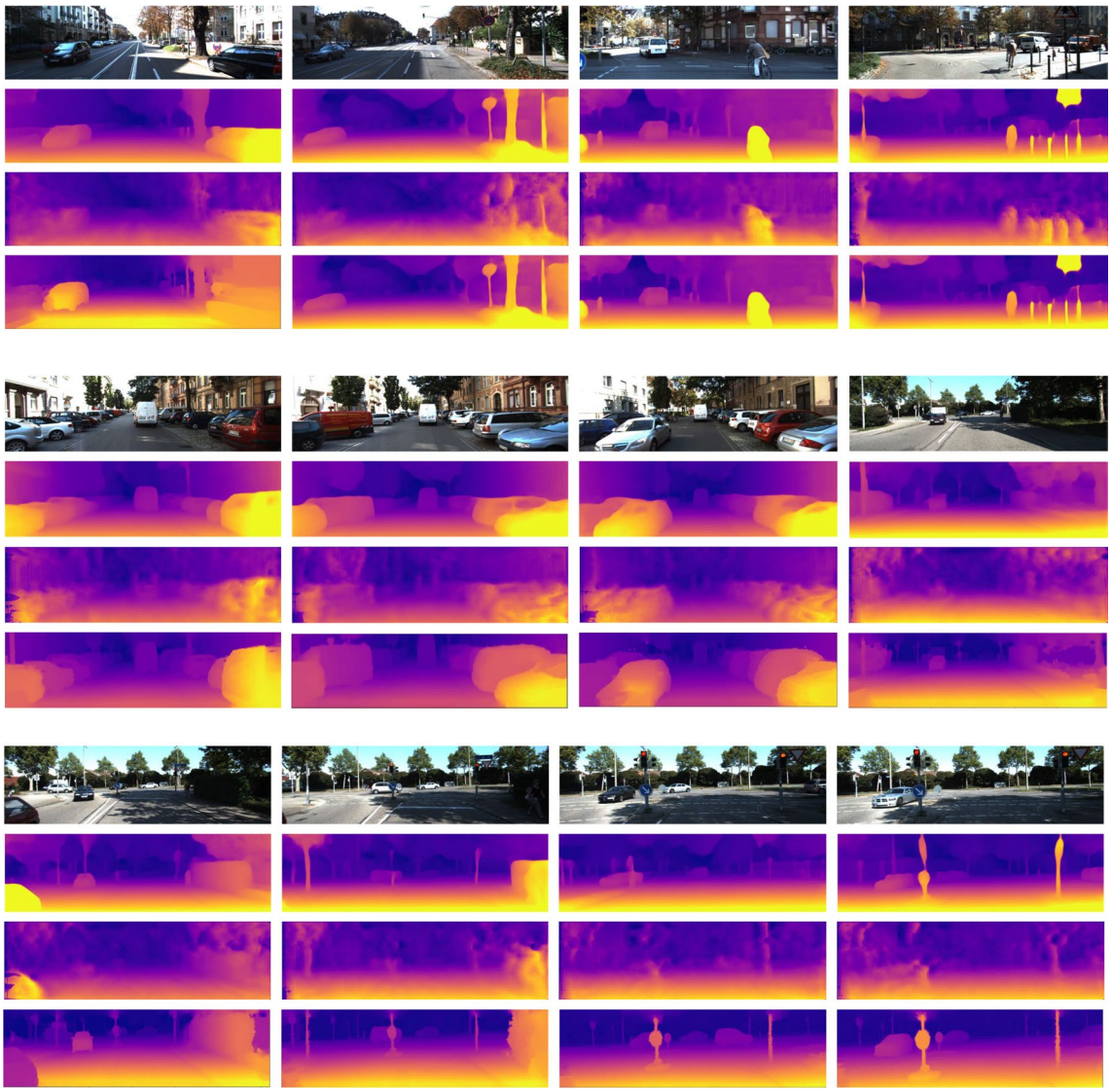

Fig. 8 This visualization contains input image (1st row), our result (2nd row), Pilzer et al. [4] approach (3rd row), and ground truth depth maps (4th row). The ground truth depth maps are interpolated from sparse LIDAR points for visualization purpose only

5 Limitation

The limitation of our approach is mainly two-fold:

1. The major limitation of GAN implementation is probably the fact that it is notorious to train. Several empirical tricks have been implemented in various literatures to make it work efficiently (i.e. in this proposed method, spectral normalization has been used in order to retain the multi-modality).
2. The proposed method is not precise enough while estimating depth from noisy images.

6 Conclusion

We have presented an efficient approach to build a GAN architecture for unsupervised depth estimation. It takes advantage of the convolution factorization for learning richer image representation along with more efficient computation. The proposed attention mechanism guides the learning of the generator's feature representations to a structured scene output. It shows significant reduction of computation is possible for deep networks without performance drop. Experiments on publicly available datasets demonstrate the efficiency of our approach and competitive performance compared to the state of the art approaches.

As a part of future work, it will be interesting to see how we can use the structured prediction based graphical models on

the disparity map to obtain better scene structures along with application of this depth information to AR and 3D vision fields.

Acknowledgements This work is partially supported by the National Science Foundation (NSF) under Grant No. 1910844.

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, pp. 2366–2374 (2014)
2. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3917–3925 (2018)
3. Godard, C., Aodha, O. M., Brostow, G. J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 270–279 (2017)
4. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 International conference on 3D vision (3DV), IEEE, pp. 587–595 (2018)
5. Mehta, S., Rastegari, M., Shapiro, L., Hajishirzi, H.: Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9190–9200 (2019)
6. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. **47**(1–3), 7–42 (2002)
7. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: learning to predict new views from the world's imagery. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5515–5524 (2016)
8. Saxena, A., Sun, M., Ng, A. Y.: Learning 3-d scene structure from a single still image. In: 2007 IEEE 11th International conference on computer vision, IEEE, pp. 1–8 (2007)
9. Konrad, J., Wang, M., Ishwar, P., Wu, C., Mukherjee, D.: Learning-based, automatic 2d-to-3d image and video conversion. IEEE Trans. Image Process. **22**(9), 3485–3496 (2013)
10. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. Int. J. Comput. Vis. **75**(1), 151–172 (2007)
11. Chen, R., Mahmood, F., Yuille, A., and Durr, N. J.: Rethinking monocular depth estimation with adversarial training. arXiv preprint. arXiv:1808.07528 (2018)
12. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision, Springer, pp. 746–760 (2012)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp. 3354–3361 (2012)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223 (2016)
15. Li, B., Shen, C., Dai, Y., Van Den Hengel, A., He, M.: 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, pp. 1119–1127 (2015)
16. Cao, Y., Wu, Z., Shen, C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans. Circuits Syst. Video Technol. **28**(11), 3174–3182 (2017)
17. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A. L.: Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2800–2809 (2015)
18. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 675–684 (2018)
19. Chen, C., Wei, J., Peng, C., Zhang, W., Qin, H.: Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion. IEEE Trans. Image Process. **29**, 4296–4307 (2020)
20. Zhan, H., Garg, R., Weerasekera, C. S., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 340–349 (2018)
21. Wang, C., Buenaposada, J. M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2022–2030 (2018)
22. Garg, R., Bg, V. K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: geometry to the rescue. In: European conference on computer vision, Springer, pp. 740–756 (2016)
23. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A. A.: Learning dense correspondence via 3d-guided cycle consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 117–126 (2016)
24. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
25. Kundu, J. N., Uppala, P. K., Pahuja, A., Babu, R. V.: Adadepth: unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2656–2665 (2018)
26. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint. arXiv:1411.1784 (2014)
27. Zhu, J.-Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232 (2017)
28. Kumar, A. C. S., Bhandarkar, S. M., Prasad, M.: Monocular depth prediction using generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 300–308 (2018)
29. Almalioglu, Y., Saputra, M. R. U., de Gusmao, P. P., Markham, A., Trigoni, N.: Ganvo: unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp. 5474–5480 (2019)
30. Hao, Z., Li, Y., You, S., and Lu, F.: Detail preserving depth estimation from a single image using attention guided networks. In: 2018 International conference on 3D vision (3DV), IEEE, pp. 304–313 (2018)
31. Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning, pp. 7354–7363, PMLR (2019)

32. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint. arXiv:1802.05957 (2018)
33. Krizhevsky, A., Sutskever, I., and Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
34. He, K., Zhang, X, Ren, S., and Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
35. Xie, S., and Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision, pp. 1395–1403 (2015)
36. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado,G. S., Davis, A., Dean, J., Devin, M. et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint. arXiv:1603.04467 (2016)
37. Kingma, D.P., and Ba, J.: Adam: a method for stochastic optimization. arXiv preprint. arXiv:1412.6980 (2014)
38. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2024–2039 (2016)
39. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(6), 1426–1440 (2019)
40. Zhou, T., Brown, M., Snavely, N., and Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1851–1858 (2017)

Sumanta Bhattacharyya received his MS degree in electrical engineering from UNC, Charlotte in 2019. His research interests include signal and image processing, computer vision, and deep learning.

Ju Shen received the Ph.D. and M.Sc. degrees in computer science from the University of Kentucky, in 2014 and the University of Birmingham, England, in 2006. He is now an assistant professor of computer science at the University of Dayton and the director of the Interactive Visual Media Lab (IVIDIA). His research interests include computer vision, machine learning, and interactive media. He has authored and co-authored more than 30 published papers. He has served as the journal editor, program chair, area chair, and TPC for a number of scientific journals and international conferences, including IEEE ICME, IEEE ISM, IEEE ICIP, IEEE BigMM, IEEE MIPR, MAICS. He is also a senior member of the IEEE society.

Stephen Welch is VP of Data Science at Mariner, where he leads a team developing deep learning-based solutions for manufacturing applications. Prior to working with Mariner, Stephen was VP of Machine Learning at Autonomous Fusion, an Atlanta-based autonomous driving startup, where Stephen lead the design, development, and deployment of machine learning algorithms for autonomous driving. Stephen has extensive experience training and deploying machine learning models across a wide variety of domains, including an on-board crash detection algorithm that is now deployed in over 1M vehicles as part of the Verizon Hum product. Stephen strives to not just develop strong technology, but to explain and communicate results in clear and accessible ways—as an adjunct professor at UNCC, Stephen teaches a 60+ person graduate level class in machine learning and computer vision. Finally, Stephen is the author of the educational YouTube channel Welch Labs, which has earned 200k+ subscribers and 10M+ views. Stephen holds 10+ US patents, and engineering degrees from Georgia Tech and UC Berkeley.

Chen Chen received the B.E. degree in automation from Beijing Forestry University, Beijing, China, in 2009 and the MS degree in electrical engineering from Mississippi State University, Starkville, in 2012 and the PhD degree in the Department of Electrical Engineering at the University of Texas at Dallas, Richardson, TX in 2016. He held a Postdoctoral Research Associate position at Center for Research in Computer Vision (CRCV), University of Central Florida, from July 2016 to June 2018. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at University of North Carolina at Charlotte. His research interests include signal and image processing, computer vision, and deep learning. He published over 50 papers in refereed journals and conferences in these areas.