Spatial Correspondence with Generative Adversarial Network: Learning Depth from Monocular Videos

Zhenyao Wu[1,*], Xinyi Wu[1,*], Xiaoping Zhang[2], Song Wang[1,3,†], Lili Ju[1,3,†]
[1]University of South Carolina, USA [2]Wuhan University, China [3]Farsee2 Technology Ltd, China
{zhenyao, xinyiw}@email.sc.edu, xpzhang.math@whu.edu.cn, songwang@cec.sc.edu, ju@math.sc.edu

Abstract

Depth estimation from monocular videos has important applications in many areas such as autonomous driving and robot navigation. It is a very challenging problem without knowing the camera pose since errors in camera-pose estimation can significantly affect the video-based depth estimation accuracy. In this paper, we present a novel SC-GAN network with end-to-end adversarial training for depth estimation from monocular videos without estimating the camera pose and pose change over time. To exploit cross-frame relations, SC-GAN includes a spatial correspondence module which uses Smolyak sparse grids to efficiently match the features across adjacent frames, and an attention mechanism to learn the importance of features in different directions. Furthermore, the generator in SC-GAN learns to estimate depth from the input frames, while the discriminator learns to distinguish between the ground-truth and estimated depth map for the reference frame. Experiments on the KITTI and Cityscapes datasets show that the proposed SC-GAN can achieve much more accurate depth maps than many existing state-of-the-art methods on monocular videos.

1. Introduction

Depth estimation from 2D images or videos is critical for many computer-vision applications, including robotics [4], autonomous driving [9, 23], 3D reconstruction [58] and augmented realities [11]. As in many other computer vision tasks, convolutional neural networks (CNNs) have been widely applied to depth estimation with significant success in recent years, such as estimating depth from single images [18, 19, 20, 40, 2, 57], stereo images [8, 38], multi-view images [32, 49, 51, 54] and monocular videos [37, 60, 59, 48, 42, 55]. Among them, depth estimation from monocular videos has drawn more and more interests in recent years since 1) it only requires one monocular camera as in many real scenarios, and 2) spatial relations between adjacent frames provide important information for depth estimation. The goal of our work focuses on developing new CNN models for depth estimation from monocular videos.

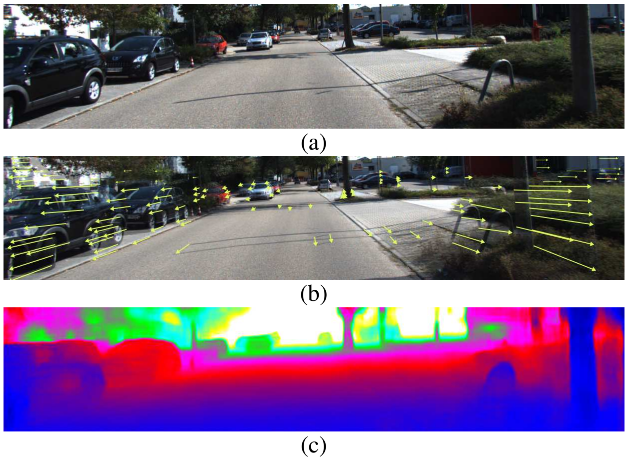
(a)
(b)
(c)

Figure 1. An illustration of the spatial relations between adjacent frames and depth estimation. (a) The reference frame; (b) the corresponding features between two adjacent frames; (c) the depth map estimated by the proposed SC-GAN.

Different from stereo matching where the input pair of stereo images are taken by two cameras with a fixed relative pose, the camera pose change between adjacent frames in videos are time-varying and unknown priorly, which makes depth estimation from monocular videos a very challenging problem. Most of the available methods address this problem by first estimating the camera pose and pose change over time, usually by training respective CNNs [46, 50, 60]. For these methods, errors in camera-pose estimation can significantly affect the accuracy of final depth estimation [51].

In this paper, we develop a novel network of SC-GAN (Spatial Correspondence with Generative Adversarial Network) to exploit latent information between adjacent frames of a monocular video and estimate the depth by supervised training. We first propose a spatial correspondence (SC) module to match the features between adjacent frames, as

[*]Equal contribution, [†]Co-corresponding authors.

shown in Figure 1. Inspired by simple observation that spatial features along different directions make different amounts of contributions in estimating the depth map, we introduce a direction-based attention (DBA) mechanism in the SC module to learn the importance of features along different directions. One key issue in building the feature relations between two frames lies in the computational and memory complexity. With large camera-pose change between frames and high image resolution, both of which are common in autonomous driving and virtual reality, the search space of corresponding features between two frames is very large. To address this issue, we down-sample patches of interest in adjacent frames using the Smolyak sparse grid method [47], which brings us both efficiency and accuracy in building cross-frame spatial relations.

In general, we employ an end-to-end adversarial training for SC-GAN, where the generator learns to estimate depth from the input frames, while the discriminator learns to distinguish between the ground-truth and estimated depth maps for the reference frame. In the experiments, we carry out a series of ablation and comparison studies on the KITTI and Cityscapes benchmark datasets, and find that the proposed SC-GAN can achieve significantly better performance than many existing state-of-the-art monocular depth estimation methods.

The major contributions of this paper are summarized below:

- We develop the SC-GAN network, with a newly designed spatial correspondence module, for depth estimation from monocular videos in an end-to-end manner.
- We make use of Smolyak sparse grids to greatly reduce the complexity of correlation calculations for the spatial correspondence of adjacent frames. As far as we know, this is the first time to use this method for solving computer vision problems.
- SC-GAN significantly promotes the state-of-the-art performance of the monocular depth estimation on the KITTI and Cityscapes datasets.

2. Related Work

The use of deep learning has promoted the depth estimation accuracies on single images [18, 19, 20, 40, 25, 52, 41] and stereo images [8, 38]. These methods are not very suitable to address the problem of depth estimation from monocular videos in that 1) without considering cross-frame relations, single-image depth estimation methods usually show limited accuracy, and 2) stereo matching assumes fixed relative pose between two input images, which is not held for adjacent frames in videos.

Depth estimation from monocular videos has attracted much interest in recent years. In [37], handcrafted features were matched between frames for depth estimation and optical flow is also used to improve the depth estimation accuracy. Zhou *et al.* [60] trained a network to estimate the relative camera pose between adjacent frames and then fed it to another network for depth estimation. DeepV2D [48] estimates the relative camera poses between a key frame and a set of nearby frames and finally generates a fused depth map on the key frame. DeepTAM [59] estimates the relative camera pose and uses it to propagate the known depth map of a key frame to other frames. Mahjourian *et al.* [42] combined camera-pose estimation and depth estimation in a single network by enforcing the 3D geometry consistency. Yin *et al.* [55] considered camera pose estimation, depth estimation and optical flow in a unified network. We can see that all these methods need to estimate camera-pose change between frames. Differently, in this paper, we directly exploit the spatial relations between adjacent frames without estimating the camera pose and pose change along the video.

Also related is another line of work on learning depth from multi-view stereo [32, 49, 51, 54]. These methods can be applied to estimate depth maps from monocular videos if treating multiple adjacent frames as multi-view images. However, many of these methods require camera poses to be given [33, 35, 54] and others need to estimate camera poses [49, 32] just like the video-based depth estimation methods mentioned above. The proposed SC-GAN does not estimate the camera poses and we show it can lead to better depth estimation from monocular videos by various experiments.

The proposed SC-GAN is a Generative Adversarial Network (GAN) [26] which has drawn broad attention in style transfer [16, 36], image-to-image translation [34, 62], image editing [61] and cross-domain image generation [5, 15]. In [45], a GAN network was proposed to refine the estimated disparity map in image-based stereo matching. In [12, 14, 1], the classical GAN was adapted to estimate the depth from a single image. In [44], the cycled generative networks are deployed to estimate depth from stereo pair in an unsupervised manner. Different from these works, we develop in this paper a new GAN network to address video-based depth estimation, which utilizes information from adjacent frames.

The sparse grid technique is an effective numerical method with high computational efficiency for representation, interpolation and integration of functions in multi-dimensional spaces, and was first proposed in [47] by Smolyak based on sparse tensor products. Since then it has been widely used in approximation theory [3], uncertainty quantification and high-dimensional integrations [6], global optimization [43], data compression [24] and etc. In this paper, we use Smolyak sparse grids for down-sampling and

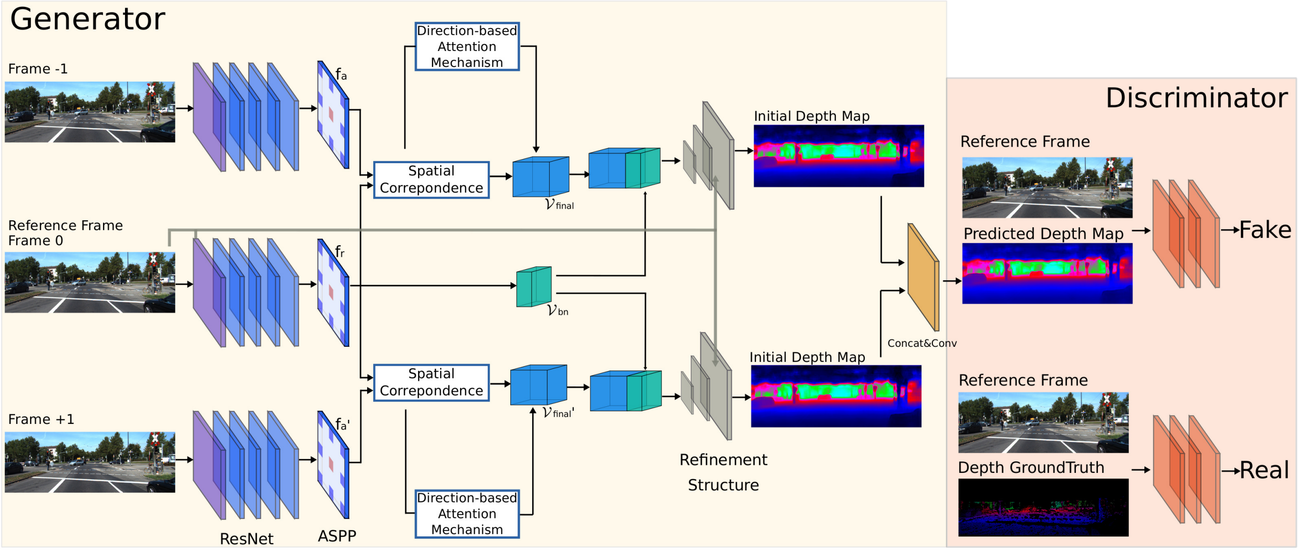

Figure 2. Architecture of the proposed SC-GAN consisting of a generator and a discriminator.

facilitating cross-frame feature correspondence in the spatial correspondence module of SC-GAN.

3. Proposed Method

The proposed SC-GAN network consists of a generator and a discriminator that compete against each other. Figure 2 presents the detailed architecture of SC-GAN. The inputs are triple adjacent frames in a video – frames -1, 0, and 1. Among them, frame 0 is the reference frame from which we seek to estimate the depth map. While this architecture can be extended to consider more or less adjacent frames, in this paper we focus on triple-frame inputs for simplicity.

3.1. Network architecture

The generator network of SC-GAN includes a spatial correspondence module, a direction-based attention mechanism and a depth map refinement module, which takes the group of triple adjacent frames as input and outputs the depth map in an end-to-end manner. Firstly, it begins by using ResNet-50 [29] to extract features from the input frames (all with the same size $W \times H$) and then for each frame the atrous spatial pyramid pooling (ASPP) [10] module is employed to extract features from multiple large receptive fields via dilated convolutional operations with dilation rates $(6, 12, 18)$. The output feature map for each frame is then a $\frac{W}{4} \times \frac{H}{4} \times \tilde{C}$ tensor where $\tilde{C}$ represents the number of channels. Note that ResNet-50 and the ASPP module are weight sharing among all three branches. Secondly, the spatial correspondence module combined with the direction-based attention mechanism is used to form correlation features for each pair of the reference frame and one of its adjacent frames. Thirdly, the correlation features and the batch normalized features of the reference frame are fed together into the weight-sharing refinement subnetwork to get the respective initial depth map for each pair. Finally, the concatenation of all initial depth maps from all pairs is the input of a 3 × 3 convolution layer to predict the final depth map of the reference frame.

During the training phase, we use the Markovian discriminator (PatchGAN [34]) which consists of 4 × 4 Convolution-InstanceNorm-LeakyReLU layers. The discriminator is used to distinguish the pair of the predicted depth map and the reference frame with the pair of the ground-truth depth map and the reference frame, and then to provide feedback to the generator.

3.2. Spatial correspondence

Inspired by Flownet [17] which introduces a "correlation layer" that performs multiplicative patch comparisons between two feature maps, we propose a spatial correspondence module to match the features. An illustration of the spatial correspondence is shown in Figure 3. Let k be the maximum displacement when corresponding the features between the feature map f_r of the reference frame and feature map f_a of one of its adjacent frames. For the feature at each position (i, j) of f_r, the search space for its corresponding feature position in f_a is a $(2k + 1) \times (2k + 1)$ square patch centered at (i, j). One common way is to define their spatial correlation features as a $\frac{W}{4} \times \frac{H}{4} \times C$ tensor $\mathcal{V}$ with $C = (2k + 1)^2$ and each of its entry is given by

$$\mathcal{V}(i, j, c) = \mathcal{C}(f_r(i, j), f_a(i + o_1, j + o_2)) \qquad (1)$$

for $o_1, o_2 \in [-k, k]$ where $c = (o_1 + k)(2k + 1) + (o_2 + k)$ and the function $\mathcal{C}$ denotes a 1×1 convolution (equivalent to the dot-product operation in this case). Thus the spatial correlation features $\mathcal{V}$ outputs C values for each position (i, j) of the reference f_r, which could be quite large when the maximum displacement k is large, especially for high-resolution input images.

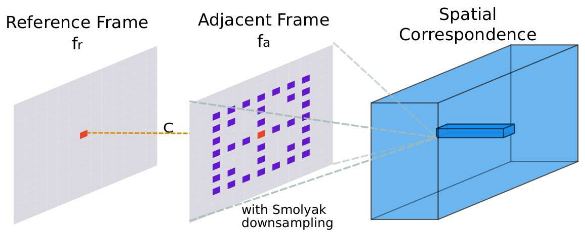

Figure 3. Architecture of the spatial correspondence module. The grey squares represent the feature maps from the reference frame and one of its adjacent frames, respectively, and the volume on the right indicates the obtained correlation features $\mathcal{V}$ defined in Eq. (1).

A typical way to reduce the search space is to down-sample the $(2k+1) \times (2k+1)$ square patch and only search for the corresponding features from a sparse set of positions. Uniform sampling is certainly a choice – as shown in Figure 4-(a), for a given sampling rate r, the third dimension (i.e., the number of channels) of the correlation feature tensor $\mathcal{V}$ can be reduced to $C = \lceil \frac{2k+1}{r} \rceil^2$ by uniform sampling. In this paper, we propose to use Smolyak sparse grids [47] for non-uniform sampling, which has been successfully used in many other applications.

The set of Smolyak sparse grids $\mathcal{S}$ in a square domain in two dimensions is defined as

$$\mathcal{S}_l = \bigcup_{\alpha_1+\alpha_2 \leqslant l} (\Theta^x_{\alpha_1} \bigotimes \Theta^y_{\alpha_2}), \tag{2}$$

where l denotes the level of the sparse grids, α_1 and α_2 are nonnegative integers, and Θ^j_α is the one-dimensional interpolation abscissas, which can be $2^\alpha + 1$ uniformly distributed points (uniform-type) or the roots (need to be accordingly scaled by the domain size) of the $(2^\alpha - 1)$-order Chebyshev polynomial $\{\cos(\frac{(n-1/2)\pi}{2^\alpha-1})\}_{n=1}^{2^\alpha-1}$ plus two end points (Chebyshev-type).

Since the distribution of Smolyak sparse grid points (especially Chebyshev-type) is highly non-uniform, we project all grid points to their nearest integer-coordinate points in the frame then remove all duplicated ones in order to avoid extra interpolations cost in practice. Such set of approximate Smolyak sparse grid-points at each level l is denoted as $\tilde{\mathcal{S}}_l$. Consequently, the third dimension of the feature tensor $\mathcal{V}$ becomes $C = |\tilde{\mathcal{S}}_l|$ with the use of $\tilde{\mathcal{S}}_l$ as the sampling points, which can be much smaller than $(2k + 1)^2$ and significantly reduces the amount of calculations while still maintaining good approximation accuracies of the correlation information. Sampled points by using two types of Smolyak sparse grids at different levels on a 49×49 square patch are illustrated in Figure 4-(b) and (c), respectively.

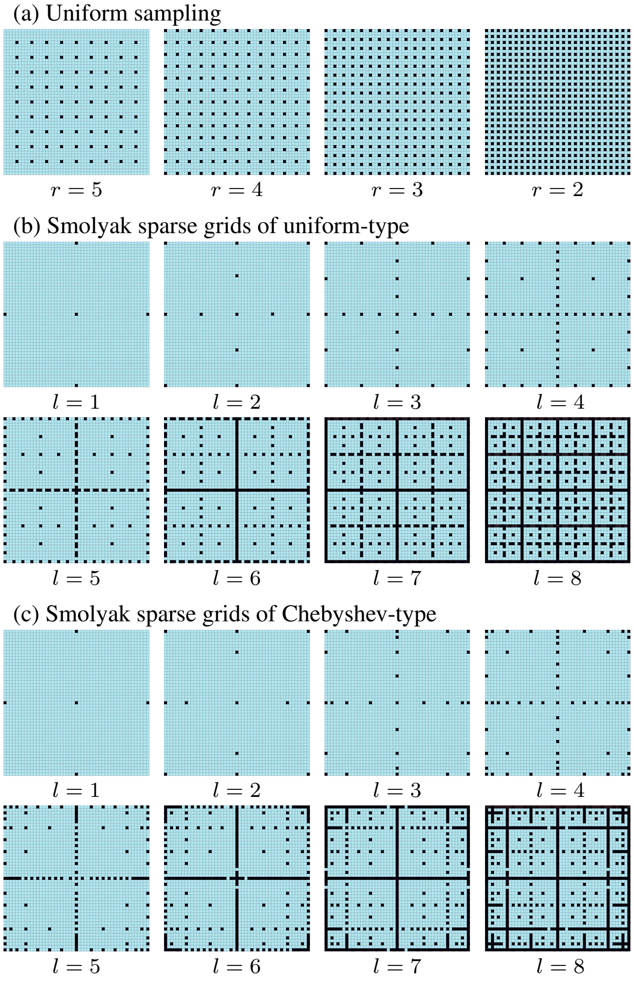

Figure 4. Sampled points by using different down-sampling methods on a 49×49 square patch. (a) Uniform samplings; (b) Smolyak sparse grids of uniform-type; (c) Smolyak sparse grids of Chebyshev-type.

In the later experiment, we will conduct ablation studies to compare the performance of using uniform down-sampling, and Smolyak sparse grids of uniform-type and Smolyak sparse grids of Chebyshev-type for down-sampling.

3.3. Direction-based attention mechanism

To enable the spatial correspondence module to selectively leverage the spatial correlation features aggregated along different directions, we propose a direction-based attention mechanism (DBA), inspired by the DSC method [31] and the Squeeze-and-Excitation Blocks [30]. An illustration of the DBA mechanism is shown in Figure 5, which consists of an adaptive average pooling layer, two fully connected (FC) layers, and a ReLU layer and generates a direction-wise attention vector for each pair of feature maps (f_r and f_a) of the reference frame and one of its adjacent frames.

The DBA mechanism starts from computing a vector $\boldsymbol{w} \in \mathbb{R}^{2\tilde{C}}$ from features of the reference and adjacent frames:

$$\boldsymbol{w} = \frac{16}{H \times W} \sum_{i=1}^{H/4} \sum_{j=1}^{W/4} \langle f_r(i,j), f_a(i,j) \rangle, \qquad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes the concatenation. Then the direction-based vector $\boldsymbol{w}_{DBA} \in \mathbb{R}^{C}$ is defined as:

$$\boldsymbol{w}_{DBA} = \boldsymbol{W}_2 \cdot \mathcal{R}(\boldsymbol{W}_1 \cdot \boldsymbol{w}), \qquad (4)$$

where $\mathcal{R}$ stands for the ReLU function and $\boldsymbol{W}_1 \in \mathbb{R}^{C \times 2\tilde{C}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{C \times C}$ are the weight matrices for two fully connected layers. The final direction-based correlation feature tensor $\mathcal{V}_{final}$ is then defined by

$$\mathcal{V}_{final}(i,j,c) = \mathcal{V}(i,j,c) \cdot \boldsymbol{w}_{DBA}(c), \qquad (5)$$

for $c = 0, 1 \cdots, C-1$.

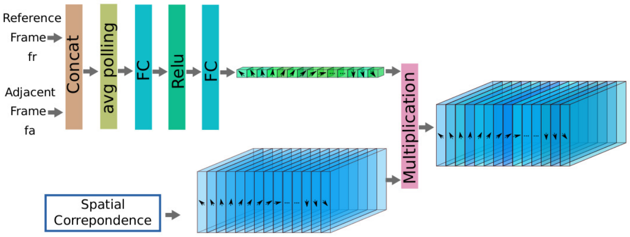

Figure 5. Architecture of the direction-based attention (DBA) mechanism.

Note that, our DBA mechanism is different from the prior DSC method [31] and the Squeeze-and-Excitation Blocks [30]. By including the pooling layer, the DBA mechanism uses global information to learn the weight of all possible directions, while the DSC method [31]) uses local information only. The Squeeze-and-Excitation Blocks [30] learns the interdependencies between channels from a single image, while the proposed DBA mechanism learns a vector to represent the weight to reflect the importance of each of the directions.

3.4. Depth map refinement

In order to provide dense depth predictions with high resolution, we build a weight-sharing refinement subnetwork to refine the predicted depth map for each group of input frames. The input of the refinement subnetwork is the concatenation of $\mathcal{V}_{final}$ and $\mathcal{V}_{bn}$, where $\mathcal{V}_{final}$ is the final correlation feature map defined in Eq. (5) and V_{bn} is the result of batch normalization on f_r.

The refinement subnetwork is depicted in Figure 6, which includes a series of deconvolution, concatenation, and convolution operations, as in [56, 17]. Following these operations, we can leverage both high-level and low-level information from three parts including the features generated from the ASPP module, features obtained after Conv0 layer in ResNet, and the original reference frame. The kernel size in convolution blocks is 3 × 3 and each deconvolution layer doubles the resolution of the result. And finally, obtain an initial estimate of the depth map with the same resolution as the original frame.

3.5. Loss function

SC-GAN contains a generator G to estimate the depth map $G(R)$ for the reference frame R as described above, and a discriminator D to distinguish the ground-truth depth map $\mathcal{M}_d$ and the predicted depth map $G(R)$ of the reference frame. Following [34, 12], SC-GAN is trained with a per-pixel loss term $\mathcal{L}_{L1}$ and an adversarial loss term $\min_G \max_D \mathcal{L}_{GAN}$:

$$\mathcal{L} = \mathcal{L}_{L1} + \lambda \min_G \max_D \mathcal{L}_{GAN}, \qquad (6)$$

where λ is the balancing factor. Since the ground-truth depth map is usually sparse, we define φ as the mask operation converting the estimated depth maps to the correspond-

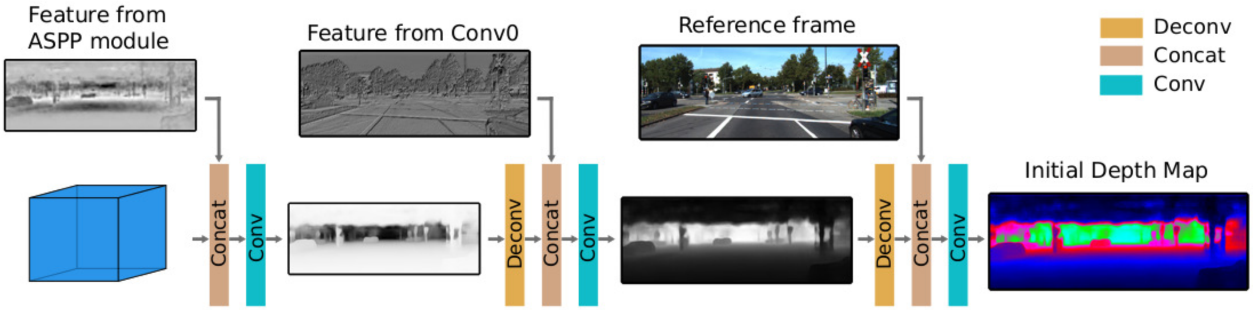

Figure 6. Architecture of the refinement subnetwork.

ing sparse ones. The per-pixel loss term $\mathcal{L}_{L1}$ is defined as:

$$\mathcal{L}_{L1} = \mathbb{E}_{\mathcal{M}_d, G(R)}[\|\mathcal{M}_d - \varphi(G(R))\|_1]. \tag{7}$$

The adversarial loss is expressed as:

$$\begin{aligned}\min_G \max_D \mathcal{L}_{GAN}(G, D) = \mathbb{E}_{R,\mathcal{M}_d}[\log D(R, \mathcal{M}_d)] + \\ \mathbb{E}_{R,G(R)}[\log(1 - D(R, \varphi(G(R))))],\end{aligned} \tag{8}$$

where G tries to minimize this loss against an adversarial D that tries to maximize it. The purpose of using the adversarial loss is to classify the overlapping pairs of the reference frame and the depth patches as being real or fake.

4. Experiments

4.1. Datasets and evaluation metrics

The following two datasets are used for performance evaluation and comparisons of the proposed SC-GAN with many existing state-of-the-art networks on depth estimation from monocular videos.

KITTI [22]: The KITTI dataset is the most commonly used benchmark in prior works for evaluating the depth, disparity and ego-motion accuracy [21, 60], which includes a full suite of data sources such as stereo videos and sparse depth maps from LIDAR. For our experiments, we only use the monocular video streams and the corresponding sparse depth maps for training and the reference frames in the test split are the same as the KITTI Raw Eigen test split [19, 60].

Cityscapes [13]: The Cityscapes dataset consists of a large set of stereo video sequences recorded in streets from 50 different cities with ground-truth disparities. Due to the focus on monocular videos, we choose the image sequences, each of which is 30-frame snippet (17Hz) around a left 8-bit image from the train, validation, and test sets (150,000 images). For each sequence, we take its left 8-bit image as the reference frame, together with its adjacent two frames, as input to the network. The ground-truth depth map for each reference frame is inferred from its disparities.

Our evaluations are based on several metrics from prior work [19] – Error metrics: absolute relative difference (Abs Rel), squared relative difference (Sq Rel), root-mean-square error (RMSE), log RMSE (RMSE log); and Accuracy metrics: the accuracy with threshold $\delta = \{1.25, 1.25^2, 1.25^3\}$ respectively. For all error metrics, the lower the better, while for the accuracy metrics, the higher the better.

4.2. Model specification

The proposed SC-GAN was implemented based on PyTorch, and all trainings were done on two Nvidia 1080 GPUs with the minibatch SGD and the Adam solver (the momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$). Following the standard approach from [34], we performed one gradient descent step on discriminator, then one step on generator. We trained our model from scratch using the training dataset with the batch size of 1, and kept the same learning rate ($lr = 0.0002$) for both generator and discriminator. We performed color normalization on the entire dataset for data preprocessing, and during the training process, all images were randomly cropped to the size of 256 × 512 and augmented with random (flip and color) transformations as done in [19]. We set the number of channel $\tilde{C} = 256$ for feature extraction, then each of the resulting feature maps was a tensor of size $64 \times 128 \times 256$. The maximum displacement k was set to be 24 in the spatial correspondence module and the maximum depth to be 80 for both datasets. For all experiments, the balance factor $\lambda = 0.1$ was used in Eq. (6). We trained SC-GAN on the KITTI dataset with 10 epochs while on the Cityscapes dataset with 12 epochs.

Table 1. Ablation study for selecting the down-sampling method for feature correspondence in SC-GAN on KITTI. The rightmost column is the computation time (measured in seconds) for processing of one group of triple frames in training.

Method	Level	C	Abs Rel	Sq Rel	RMSE	Time
Uniform sampling	$r = 4$	169	0.071	0.227	2.444	0.24
	$r = 3$	289	0.068	0.207	2.263	0.26
	$r = 2$	625	0.064	0.181	2.116	0.39
Smolyak sparse grids-Uniform	$l = 5$	145	0.073	0.228	2.316	0.23
	$l = 6$	289	0.069	0.209	2.312	0.26
	$l = 7$	481	0.066	0.201	2.270	0.32
	$l = 8$	737	0.063	0.181	2.084	0.41
Smolyak sparse grids-Chebyshev	$l = 5$	129	0.074	0.220	2.319	0.23
	$l = 6$	261	0.069	0.212	2.464	0.26
	$l = 7$	441	0.063	0.178	2.129	0.32
	$l = 8$	669	0.062	0.174	2.082	0.39

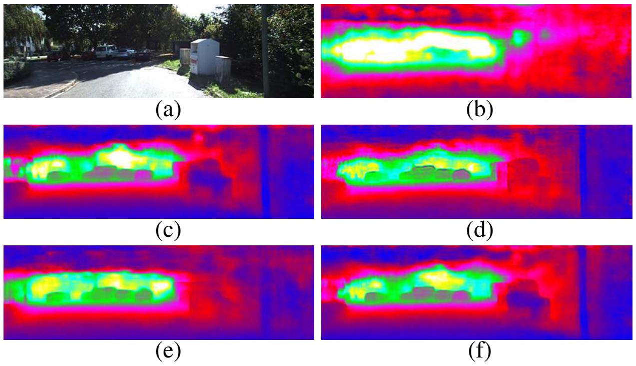

Figure 7. Depth maps resulting from different model variants of SC-GAN: (a) the reference frame; (b) the depth map estimated without the spatial correspondence module; (c) the depth map estimated without the direction-based attention module; (d) the depth map estimated without the refinement subnetwork; (e) the depth map estimated without the adversarial loss; (f) the depth map estimated by the full version of SC-GAN.

Table 2. Comparison of a number of different model variants of SC-GAN on KITTI.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
w/o Spatial correspondence	0.079	0.222	2.301	0.111	0.949	0.981	0.994
w/o DBA mechanism	0.065	0.174	2.182	0.094	0.956	0.991	0.997
w/o Refinement	0.103	0.409	3.354	0.150	0.895	0.979	0.991
w/o Adversarial loss	0.069	0.234	2.653	0.120	0.934	0.990	0.995
Full version of SC-GAN	0.063	0.178	2.129	0.097	0.961	0.993	0.998

Table 3. Performance comparison of SC-GAN and some existing state-of-the-art networks on KITTI. Note that the $\star$ marks the method is in the semi-supervised or unsupervised manner.

Method	Input	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [19]	Single image	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu *et al.* [40]	Single image	0.202	1.614	6.523	0.275	0.678	0.895	0.965
DORN [20]	Single image	0.072	0.307	2.727	0.120	0.932	0.984	0.994
DfUSMC [28]*	Stereo	0.346	5.984	8.879	0.454	0.617	0.796	0.874
Godard *et al.* [25]*	Stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Kuznietsov *et al.* [39]	Stereo	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Guo *et al.* [27]	Stereo	0.097	0.653	4.170	0.170	0.889	0.967	0.986
Zhou *et al.* [60]*	Video	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yin *et al.* [55]*	Video	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Yang *et al.* [53]	Video	0.097	0.734	4.442	0.187	0.888	0.958	0.980
Teed *et al.* [48]	Video	0.091	0.582	3.644	0.154	0.923	0.970	0.987
SC-GAN	Video	**0.063**	**0.178**	**2.129**	**0.097**	**0.961**	**0.993**	**0.998**

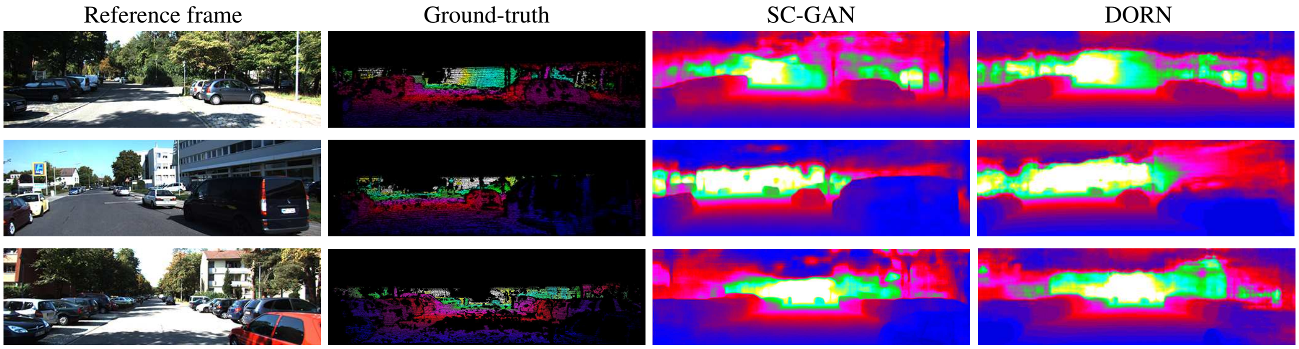

Figure 8. Visualization of three testing results from KITTI. From left to right: the reference frame, the ground-truth depth map, the predicted depth map by SC-GAN and the predicted depth map by DORN.

4.3. Ablation studies

We first carry out a study to select the down-sampling method and parameters for feature correspondence in the proposed SC-GAN. Without down-sampling, we need search all $C = 49 \times 49 = 2,401$ points in the patch. The uniform sampling with different rates ($r = 4, 3, 2$) and the approximated uniform-type and Chebyshev-type Smolyak sparse grids $\tilde{\mathcal{S}}_l$ at different levels ($l = 5, 6, 7, 8$) are tested and compared on the KITTI dataset. The results in Table 1 show that: 1) compared to uniform sampling, the Smolyak sparse grids tend to be more efficient and effective for spatial correlation when C becomes larger since uniform sampling may contain a lot of redundant information; 2) by seeking a compromise between the time efficiency (and memory cost) and the depth estimation accuracy, we choose the Chebyshev-type Smolyak sparse grid at level 7, which clearly beats the other two down-sampling methods with similar numbers of sampled points, i.e., the uniform sampling with $r = 2$ and the uniform-type Smolyak sparse grid at level 7.

In all the remaining experiments, the Chebyshev-type $\tilde{\mathcal{S}}_7$ (441 points, about 18.37% of the original 2,401 points) is used as the down-sampling method in SC-GAN. In this case, it took about 50 hours of training for SC-GAN on the KITTI dataset and 40 hours on the Cityscapes dataset.

Table 4. Performance comparison of SC-GAN and some state-of-the-art networks, when trained on Cityscapes and then tested on KITTI.

Method	Input	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [19]	Single Image	0.423	4.373	8.487	0.356	0.655	0.871	0.951
Godard *et al.* [25]	Stereo	0.233	3.533	7.412	0.292	0.700	0.892	0.953
Caser *et al.* [7]	Video	0.153	1.109	5.557	0.227	0.796	0.934	0.975
SC-GAN	Video	**0.149**	**0.921**	**4.812**	**0.192**	**0.818**	**0.954**	**0.987**

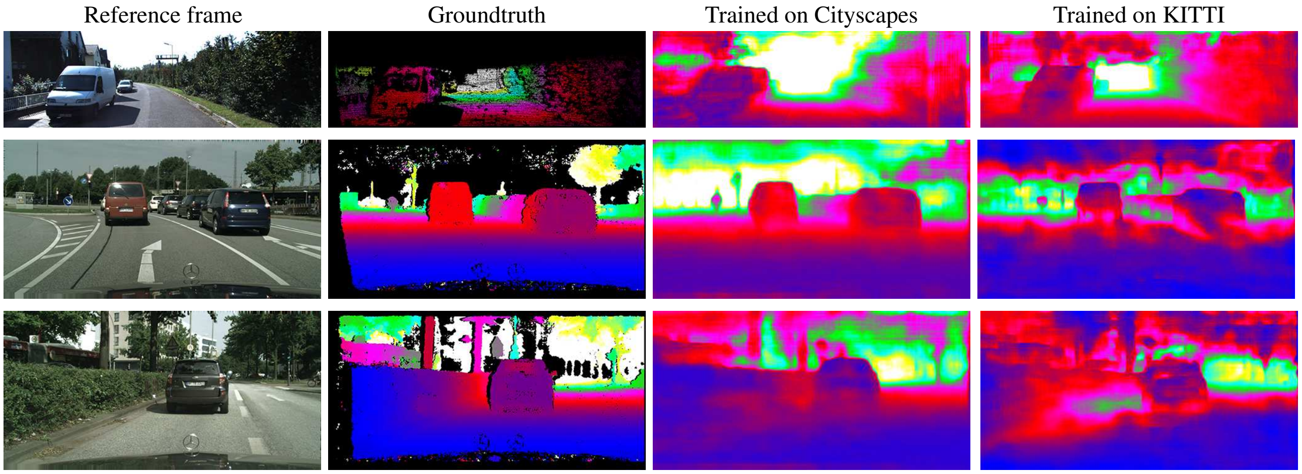

Figure 9. Three sample results in testing the generalization ability of SC-GAN. First row is an example from KITTI, and the bottom two rows are examples from Cityscapes. From left to right are the reference frame, the ground-truth depth map, the depth map predicted by SC-GAN trained on Cityscapes, the depth map predicted by SC-GAN trained on KITTI.

Next, we conduct ablation study to justify different modules of SC-GAN on the KITTI dataset, including 1) the spatial correspondence module; 2) the direction-based attention mechanism; 3) the refinement subnetwork; and 4) the adversarial loss. The quantitative results are reported in Table 2 and a sample result of depth maps are shown in Figure 7. We can see that all these four modules can help improve the depth estimation.

4.4. Comparisons with existing networks

Firstly, we evaluate and compare the performance of SC-GAN with eleven existing state-of-the-art networks [19, 40, 20, 28, 25, 39, 27, 60, 55, 53, 48] for depth estimation on the KITTI dataset. All these networks are trained on the KITTI dataset and Table 3 reports the evaluation results. It is easy to see that SC-GAN achieves the best performance (with significant better results) under all error and accuracy metrics. Figure 8 presents three examples of the depth maps estimated by SC-GAN and DORN [20].

We also evaluate the *generalization ability* of SC-GAN. In this case, SC-GAN is trained only on the Cityscapes dataset and then tested on the KITTI dataset. Table 4 reports the corresponding performance evaluation results, from which we see that SC-GAN again significantly outperforms the three comparison methods [19, 25, 7]. The results of this generalization test for other comparison methods are not available in literature. We also note that when testing on the KITTI dataset, SC-GAN trained on CitysCities paces can even get comparable performance to several supervised and semi-supervised models [25, 55, 39] that are trained on KITTI (see Table 3 and Table 4). Figure 9 visually illustrates the estimated depth maps for three examples produced by SC-GAN that are trained on different datasets. All these results clearly demonstrate the excellent generalization ability of the proposed SC-GAN.

5. Conclusion

In this paper, we developed a novel end-to-end SC-GAN network for depth estimation from monocular videos. SC-GAN consists of a generator and a discriminator. In the generator, a spatial correspondence module is designed to match the features between the reference frame and its adjacent frames. We proposed to use the approximate Smolyak sparse grids for patch down-sampling that can significantly speed up the feature correspondence. We further developed a direction-based attention mechanism to learn the importance of features in different directions, and included a refinement subnetwork to refine the initially estimated depth maps. Extensive experiments on the KITTI and Cityscapes datasets demonstrate that the proposed SC-GAN significantly promotes the state-of-the-art performance of the depth estimation from monocular videos.

References

[1] Filippo Aleotti, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In *European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2810, 2018.

[3] Volker Barthelmann, Erich Novak, and Klaus Ritter. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, 12(4):273–288, 2000.

[4] Joydeep Biswas and Manuela Veloso. Depth camera based localization and navigation for indoor mobile robots. In *RGB-D Workshop at RSS*, volume 2011, page 21, 2011.

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta numerica*, 13:147–269, 2004.

[7] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. *arXiv preprint arXiv:1811.06152*, 2018.

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2722–2730, 2015.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Qiuyu Chen, Ryoma Bise, Lin Gu, Yinqiang Zheng, Imari Sato, Jenq-Neng Hwang, Sadakazu Aiso, and Nobuaki Imanishi. Virtual blood vessels in complex background using stereo x-ray images. In *IEEE International Conference on Computer Vision Workshops*, pages 99–106, 2017.

[12] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[14] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Monocular depth prediction using generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–308, 2018.

[15] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[18] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015.

[19] David Eigen, Christian Puhrsch, and Rob 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.

[20] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018.

[21] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)*, pages 740–756. Springer, 2016.

[22] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[23] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012.

[24] Thomas Gerstner. Multiresolution visualization and compression of global topographic data. *GeoInformatica*, 7(1):7–32, 2003.

[25] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[27] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *European Conference on Computer Vision (ECCV)*, pages 484–500, 2018.

[28] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5413–5421, 2016.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[31] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations*, 2019.

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[35] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.

[37] Kevin Karsch, Ce Liu, and S Kang. Depth extraction from video using non-parametric sampling-supplemental material. In *European Conference on Computer Vision (ECCV)*. Citeseer, 2012.

[38] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[39] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[40] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2016.

[41] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 155–163, 2018.

[42] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5667–5675, 2018.

[43] Erich Novak and Klaus Ritter. Global optimization using hyperbolic cross points. In *State of the Art in global Optimization*, pages 19–33. Springer, 1996.

[44] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe. Unsupervised adversarial depth estimation using cycled generative networks. In *International Conference on 3D Vision (3DV)*, pages 587–595. IEEE, 2018.

[45] Can Pu, Runzi Song, Nanbo Li, and Robert B Fisher. Sdf-gan: Semi-supervised depth fusion with multi-scale adversarial networks. *arXiv preprint arXiv:1803.06657*, 2018.

[46] Danilo Jimenez Rezende, SM Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *Advances in Neural Information Processing Systems*, pages 4996–5004, 2016.

[47] Sergei Abramovich Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Doklady Akademii Nauk*, volume 148, pages 1042–1045. Russian Academy of Sciences, 1963.

[48] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018.

[49] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[50] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2022–2030, 2018.

[51] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018.

[52] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, pages 842–857. Springer, 2016.

[53] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[54] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2018.

[55] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1983–1992, 2018.

[56] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level

feature learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2018–2025. IEEE, 2011.

[57] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[58] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[59] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *European Conference on Computer Vision (ECCV)*, pages 822–838, 2018.

[60] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 7, 2017.

[61] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision (ECCV)*, pages 597–613. Springer, 2016.

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.