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

Zhenyao Wu¹,∗, Xinyi Wu¹,∗, Xiaoping Zhang², Song Wang¹,³,†, Lili Ju¹,³,†
¹University of South Carolina, USA ²Wuhan University, China ³Farsee2 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, 53], 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.

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, 23, 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 [27], 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
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 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) 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 $W_4 \times H_4 \times \tilde{C}$ 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 $3 \times 3$ convolution layer to predict the final depth map of the reference frame.

During the training phase, we use the Markovian discriminator (PatchGAN) which consists of $4 \times 4$ 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 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 $W_4 \times H_4 \times C$ tensor $\mathcal{V}$ with $C = (2k + 1)^2$ and each of its entry is given by

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

for $o_1, o_2 \in [-k, k]$ where $c = (o_1 + k)(2k + 1) + (o_2 + k)$ and the function $C$ denotes a $1 \times 1$ convolution (equivalent to the dot-product operation in this case). Thus the spatial
correlation features $V$ 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.

![Diagram](image-url)

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 $V$ defined in Eq. [1].

A typical way to reduce the search space is to downsample 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 [3](a), for a given sampling rate $r$, the third dimension (i.e., the number of channels) of the correlation feature tensor $V$ can be reduced to $C = \lceil \frac{(2k+1)^2}{r^2} \rceil$ 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 $S$ in a square domain in two dimensions is defined as

$$S_l = \bigcup_{\alpha_1 + \alpha_2 \leq l} \Theta_{\alpha_1}^x \Theta_{\alpha_2}^y,$$

where $l$ denotes the level of the sparse grids, $\alpha_1$ and $\alpha_2$ are nonnegative integers, and $\Theta_{\alpha}^x$ is the one-dimensional interpolation abscissas, which can be $2^n + 1$ uniformly distributed points (uniform-type) or the roots (need to be accordingly scaled by the domain size) of the $(2^n - 1)$-order Chebyshev polynomial \{$\cos\left(\frac{(n-1/2)\pi}{2^{n-1}}\right)$\}$^{2^n-1}$ 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 $S_l$. Consequently, the third dimension of the feature tensor $V$ becomes $C = |S_l|$ with the use of $S_l$ 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 $49 \times 49$ square patch are illustrated in Figure [4](b) and (c), respectively.

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 \(w \in \mathbb{R}^{2C}\) from features of the reference and adjacent frames:

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

where \(\langle \cdot, \cdot \rangle\) denotes the concatenation. Then the direction-based vector \(w_{DBA} \in \mathbb{R}^{C}\) is defined as:

\[
w_{DBA} = W_2 \cdot \mathcal{R}(W_1 \cdot w),
\]

where \(\mathcal{R}\) stands for the ReLU function and \(W_1 \in \mathbb{R}^{C \times 2C}\) and \(W_2 \in \mathbb{R}^{C \times C}\) are the weight matrices for two fully connected layers. The final direction-based correlation feature tensor \(V_{final}\) is then defined by

\[
V_{final}(i,j,c) = V(i,j,c) \cdot w_{DBA}(c),
\]

for \(c = 0, 1, \ldots, C - 1\).

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 \(V_{final}\) and \(V_{bn}\), where \(V_{final}\) 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 \(3 \times 3\) 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 \(M_d\) and the predicted depth map \(G(R)\) of the reference frame. Following [34, 12], SC-GAN is trained with a per-pixel loss term \(L_{L1}\) and an adversarial loss term \(\min_G \max_D L_{GAN}\):

\[
\mathcal{L} = L_{L1} + \lambda \min_G \max_D L_{GAN},
\]

where \(\lambda\) is the balancing factor. Since the ground-truth depth map is usually sparse, we define \(\varphi\) as the mask operation converting the estimated depth maps to the correspond-
ing sparse ones. The per-pixel loss term $L_{L1}$ is defined as:

$$L_{L1} = \mathbb{E}_{M_d, G(R)}[||M_d - \varphi(G(R))||_1].$$

The adversarial loss is expressed as:

$$\min_G \max_D L_{GAN}(G, D) = \mathbb{E}_{G, R, M_d}[\log D(R, M_d)] + \mathbb{E}_{R, G(R)}[\log(1 - D(R, \varphi(G(R))))],$$

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 $256 \times 512$ and augmented with random (flip and color) transformations as done in [19]. We set the number of channel $C = 256$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Level</th>
<th>$C$</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>RMSE</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform sampling</td>
<td>r = 4</td>
<td>169</td>
<td>0.071</td>
<td>0.227</td>
<td>2.444</td>
<td>0.24</td>
</tr>
<tr>
<td>Smolyak sparse grids-</td>
<td>l = 5</td>
<td>145</td>
<td>0.073</td>
<td>0.228</td>
<td>2.316</td>
<td>0.23</td>
</tr>
<tr>
<td>Smolyak sparse grids-</td>
<td>l = 6</td>
<td>289</td>
<td>0.068</td>
<td>0.207</td>
<td>2.263</td>
<td>0.26</td>
</tr>
<tr>
<td>Smolyak sparse grids-</td>
<td>l = 7</td>
<td>625</td>
<td>0.064</td>
<td>0.181</td>
<td>2.116</td>
<td>0.39</td>
</tr>
<tr>
<td>Smolyak sparse grids-</td>
<td>l = 8</td>
<td>737</td>
<td>0.066</td>
<td>0.201</td>
<td>2.270</td>
<td>0.32</td>
</tr>
<tr>
<td>Smolyak sparse grids-</td>
<td>l = 9</td>
<td>873</td>
<td>0.063</td>
<td>0.181</td>
<td>2.084</td>
<td>0.41</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>δ &lt; 1.25</th>
<th>δ &lt; 1.25²</th>
<th>δ &lt; 1.25³</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Spatial correspondence</td>
<td>0.079</td>
<td>0.222</td>
<td>2.301</td>
<td>0.111</td>
<td>0.949</td>
<td>0.981</td>
<td>0.994</td>
</tr>
<tr>
<td>w/o DBA mechanism</td>
<td>0.065</td>
<td>0.174</td>
<td>2.182</td>
<td>0.094</td>
<td>0.956</td>
<td>0.991</td>
<td>0.997</td>
</tr>
<tr>
<td>w/o Refinement</td>
<td>0.103</td>
<td>0.409</td>
<td>3.354</td>
<td>0.150</td>
<td>0.895</td>
<td>0.979</td>
<td>0.991</td>
</tr>
<tr>
<td>w/o Adversarial loss</td>
<td>0.069</td>
<td>0.234</td>
<td>2.653</td>
<td>0.120</td>
<td>0.934</td>
<td>0.990</td>
<td>0.995</td>
</tr>
<tr>
<td>Full version of SC-GAN</td>
<td>0.063</td>
<td>0.178</td>
<td>2.129</td>
<td>0.097</td>
<td>0.961</td>
<td>0.993</td>
<td>0.998</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>δ &lt; 1.25</th>
<th>δ &lt; 1.25²</th>
<th>δ &lt; 1.25³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigen et al. [19]</td>
<td>Single</td>
<td>0.203</td>
<td>1.548</td>
<td>6.307</td>
<td>0.282</td>
<td>0.702</td>
<td>0.890</td>
<td>0.958</td>
</tr>
<tr>
<td>Liu et al. [40]</td>
<td>Single</td>
<td>0.202</td>
<td>1.614</td>
<td>6.523</td>
<td>0.275</td>
<td>0.678</td>
<td>0.895</td>
<td>0.965</td>
</tr>
<tr>
<td>DORN [20]</td>
<td>Single</td>
<td>0.072</td>
<td>0.307</td>
<td>2.727</td>
<td>0.120</td>
<td>0.932</td>
<td>0.984</td>
<td>0.994</td>
</tr>
<tr>
<td>DIUSMC * [28]</td>
<td>Stereo</td>
<td>0.346</td>
<td>5.984</td>
<td>8.879</td>
<td>0.454</td>
<td>0.617</td>
<td>0.796</td>
<td>0.874</td>
</tr>
<tr>
<td>Godard et al. [25]</td>
<td>Stereo</td>
<td>0.148</td>
<td>1.344</td>
<td>5.927</td>
<td>0.247</td>
<td>0.803</td>
<td>0.922</td>
<td>0.964</td>
</tr>
<tr>
<td>Kuznietsov et al. [39]</td>
<td>Stereo</td>
<td>0.113</td>
<td>0.741</td>
<td>4.621</td>
<td>0.189</td>
<td>0.862</td>
<td>0.960</td>
<td>0.986</td>
</tr>
<tr>
<td>Guo et al. [27]</td>
<td>Stereo</td>
<td>0.097</td>
<td>0.653</td>
<td>4.170</td>
<td>0.170</td>
<td>0.889</td>
<td>0.967</td>
<td>0.986</td>
</tr>
<tr>
<td>Zhou et al. [60]</td>
<td>Video</td>
<td>0.208</td>
<td>1.768</td>
<td>6.856</td>
<td>0.283</td>
<td>0.678</td>
<td>0.885</td>
<td>0.957</td>
</tr>
<tr>
<td>Yin et al. [55]</td>
<td>Video</td>
<td>0.155</td>
<td>1.296</td>
<td>5.857</td>
<td>0.233</td>
<td>0.793</td>
<td>0.931</td>
<td>0.973</td>
</tr>
<tr>
<td>Yang et al. [53]</td>
<td>Video</td>
<td>0.097</td>
<td>0.734</td>
<td>4.442</td>
<td>0.187</td>
<td>0.888</td>
<td>0.958</td>
<td>0.980</td>
</tr>
<tr>
<td>Teed et al. [48]</td>
<td>Video</td>
<td>0.091</td>
<td>0.582</td>
<td>3.644</td>
<td>0.154</td>
<td>0.923</td>
<td>0.970</td>
<td>0.987</td>
</tr>
<tr>
<td>SC-GAN</td>
<td>Video</td>
<td>0.063</td>
<td>0.178</td>
<td>2.129</td>
<td>0.097</td>
<td>0.961</td>
<td>0.993</td>
<td>0.998</td>
</tr>
</tbody>
</table>

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 $\tilde{S}_l$ 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 $\tilde{S}_7$ (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.

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

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, 20, 25, 28, 39, 37, 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 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


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


