

Sparse Auxiliary Networks for Unified Monocular Depth Prediction and Completion

Vitor Guizilini Rareş Ambruş Wolfram Burgard Adrien Gaidon
Toyota Research Institute (TRI), Los Altos, CA
{first.lastname}@tri.global

Abstract

Estimating scene geometry from data obtained with cost-effective sensors is key for robots and self-driving cars. In this paper, we study the problem of predicting dense depth from a single RGB image (monodepth) with optional sparse measurements from low-cost active depth sensors. We introduce Sparse Auxiliary Networks (SANs), a new module enabling monodepth networks to perform both the tasks of depth prediction and completion, depending on whether only RGB images or also sparse point clouds are available at inference time. First, we decouple the image and depth map encoding stages using sparse convolutions to process only the valid depth map pixels. Second, we inject this information, when available, into the skip connections of the depth prediction network, augmenting its features. Through extensive experimental analysis on one indoor (NYUv2) and two outdoor (KITTI and DDAD) benchmarks, we demonstrate that our proposed SAN architecture is able to simultaneously learn both tasks, while achieving a new state of the art in depth prediction by a significant margin.

1. Introduction

Dense scene geometry can be directly measured using active sensors (e.g., LiDAR, structured light) or estimated from RGB cameras (e.g., via stereo matching, structure from motion, monocular depth networks). Both approaches have complementary strengths and failure modes (e.g., rain or low light). Consequently, a robust perception system must leverage both modalities while still retaining functionality when only one is available. In this paper, we propose a learning algorithm and model that can satisfy these desiderata with a simple sensor suite: a single monocular RGB camera combined with any low-cost active depth sensor returning only a few 3D points per scene.

Monocular depth prediction is becoming a cornerstone capability for a wide range of robotic applications where RGB cameras are ubiquitous [23, 49, 51]. Recently, self-

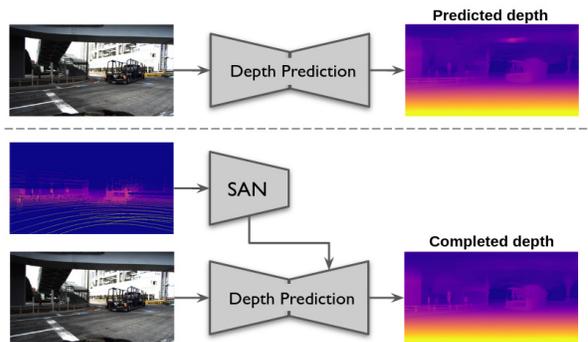


Figure 1: **Our proposed joint task learning SAN architecture** produces state of the art monocular depth estimates from a single image (prediction), which can be further improved by also providing a sparse depth map (completion) **without changing the model**.

supervised methods trained only on raw videos demonstrated that robots with a single camera can learn and predict dense depth information [2, 14, 15, 17, 58, 60], especially as the quantity of data increases [18]. However, in practice an active range sensor is often available, and can be used to either provide further supervision at training time [7, 9, 10, 30, 53] or also during inference [24, 33, 39], in a task known as depth completion. Even though sparse, recent works [20] have shown that even a few pixels containing valid depth information is enough to boost performance, and therefore should not be discarded. Importantly, these two tasks, depth *prediction* and *completion*, are treated as separate problems with different architectures. No method to date tackles the issue of *using all the information available from both modalities at both training and inference time*, including if only partially available (e.g., due to sensor blackout, occlusion, or environmental conditions).

Our **main contribution** is a novel architecture, *Sparse Auxiliary Networks* (SANs, cf. Fig. 2), that enables a monocular depth *prediction* network to also perform depth *completion* in the presence of optional sparse 3D measurements at inference time. Note that the *same architecture*

and weights can dynamically perform either task at inference time, depending on the presence or not of sparse depth measurements. Our model relies on a sparse depth convolutional encoder to inject depth information, when available, into the skip connections of state-of-the-art encoder-decoder networks for depth prediction. Our **second contribution** is a thorough experimental evaluation on three challenging outdoor (KITTI [12] and DDAD [18]) and indoor (NYUv2 [40]) datasets, demonstrating that our SAN architecture *boosts monocular depth prediction performance and sets a new state of the art in this task*.

2. Related Work

2.1. Depth Prediction

Monocular depth prediction has been gaining in popularity in the robotics community, with methods generally falling into different categories depending on the data used to derive the learning signal. *Self-supervised learning* methods aim to predict depth directly from monocular images, by imposing a photometric loss on temporally adjacent frames [59] or on corresponding stereo images [14]. Owing to its simplicity and wide availability of raw data, a wide range of body of work has addressed this topic, combining it with optical flow [54, 57], uncertainty estimation [37], semantic segmentation [19, 44], instance segmentation [2], keypoint estimation [41] and visual odometry [51, 52].

By contrast, *supervised learning* methods apply a regression loss using ground truth depth supervision, either by minimizing mean squared error [7] or through ordinal regression [9]. In addition to the standard regression loss, methods additionally use planar patches as guidance [30], impose 3D geometric constraints [53], use surface normals as regularization [38, 50], exploit task consistency constraints between depth, normals and semantic segmentation [56], or use semantic guidance [26, 35]. A number of methods use Structure-from-Motion [27] or distil stereo information [21] to use as supervisory signal during training. Our method is conceptually similar to [19], where the authors use pixel-adaptive convolutions to distill features from a semantic segmentation network. Instead, we propose novel *Sparse Residual Blocks* which leverage Minkowski convolutions [4] and are specifically designed to account for the sparse nature of our supervisory signal.

2.2. Depth Completion

While a high number of methods exists that focus purely on depth data, ranging from bilateral filters [43] to recent CNN densification methods [45], we will focus on methods that rely on RGB images as additional information. In the case when the depth signal is sparse (e.g., LiDAR), methods typically rely on RGB-based appearance as guidance and additionally devise custom convolutions and propagate con-

fidence to consecutive layers [8], or use content-dependent and spatially-variant guiding convolutions [42]. Alternative sources of information including confidence masks and object cues [47] as well as exploiting cross-attention between the RGB and depth encoders [31] can also be used.

To avoid depth mixing typically induced by the standard MSE loss a binned depth representation trained using a cross-entropy loss has been shown to work [24]. When additional temporally-adjacent frames are available a proxy photometric loss can be derived to further constrain densification [33, 55], while in this setting as little as 4 LiDAR beams are enough to provide a meaningful supervisory signal [20]. Note that our proposed method does not explicitly model any relationship between the two input modalities (RGB and depth), but rather learns these at a feature level.

3. Methodology

3.1. Monocular Depth Estimation

Prediction. The aim of monocular depth prediction is to learn a function $f_P : I \rightarrow D$ that takes as input image I and recovers a *predicted* depth $\hat{D}_P = f_P(I(p))$ for every pixel $p \in I$ (i.e., a dense depth map). In a supervised setting, we have access to sparse ground truth depth D at training time, as acquired by an independent sensor and projected back onto the camera’s image plane. Thus we treat monocular depth estimation purely as a regression problem, and learn an estimator f_P parameterized by $\theta_P = \{\theta_I\}$ by solving:

$$\hat{\theta}_P = \operatorname{argmin}_{\theta_P} \mathcal{L}_{sup}(f_P(I; \theta_P), D). \quad (1)$$

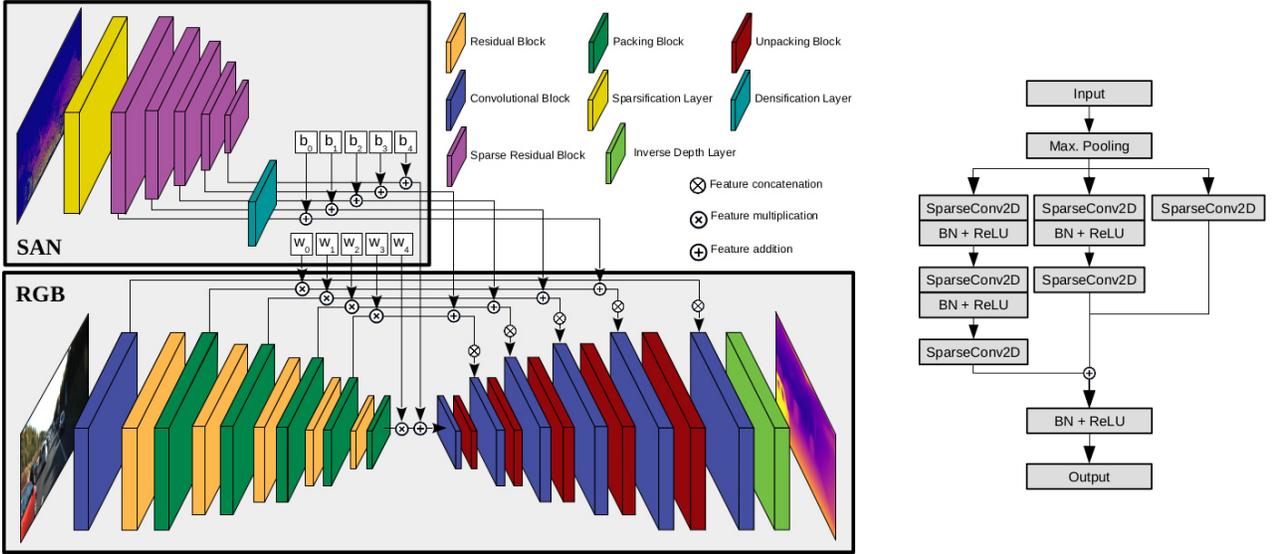
Completion. In the monocular depth completion task, we also have access to sparse ground truth depth \tilde{D} during inference (usually a subset of D [40], or collected by noisier/sparser sensors [11]). This information can be used in conjunction with I to generate a *completed* dense depth map $\hat{D}_C = f_C(I(p), \tilde{D}(p))$, where f_C is an estimator parameterized by $\theta_C = \{\theta_I, \theta_D\}$ learned by solving:

$$\hat{\theta}_C = \operatorname{argmin}_{\theta_C} \mathcal{L}_{sup}(f_C(I, \tilde{D}; \theta_C), D). \quad (2)$$

Note that f_C contains f_P , in the sense that it uses the same parameters θ_I to process the input image I , while incorporating θ_D to process the input depth map \tilde{D} . This design choice is one of the core insights of this paper, as it enables feature sharing between the tasks of depth prediction and completion (see Fig. 2).

Training Loss. Our supervised objective is the *Scale-Invariant Logarithmic* loss (SILog) [7], composed by the sum of the variance and the weighted squared mean of the error in log space $\Delta d = \log d - \log \hat{d}$:

$$\mathcal{L}(D, \hat{D}) = \frac{1}{N} \sum_{d \in D} \Delta d^2 - \frac{\lambda}{N^2} \left(\sum_{d \in D} \Delta d \right)^2, \quad (3)$$



(a) **Proposed PackNet-SAN architecture.** Residual, packing, unpacking and convolutional blocks, as well as the inverse depth layer, are detailed in [18]. The sparse residual block is detailed in Fig. 2b, the sparsification and densification layers are defined in Eqs. 6 and 7, and w and b are learnable parameters defined in Eq. 8.

(b) **Sparse residual block (SRB).** Each *SparseConv2D* layer is a Minkowski 2D Convolution [4], *BN* is Batch Normalization [25] and *ReLU* are Rectified Linear Units [13].

Figure 2: **Our proposed SAN architecture for the joint learning of monocular depth prediction and completion**, using PackNet [18] as the depth prediction network (best visualized in color).

where N is the number of valid pixels in D (invalid pixels are masked out and not considered during optimization). The coefficient λ determines the emphasis in minimizing the variance of the error. Following previous works [30], we use $\lambda = 0.85$ in all experiments. In order to train both tasks simultaneously, we add the losses generated by both output depth maps relative to the same ground truth, so that

$$\mathcal{L}(D, \hat{D}_P, \hat{D}_C) = \mathcal{L}(D, \hat{D}_P) + \mathcal{L}(D, \hat{D}_C). \quad (4)$$

3.2. Sparse Auxiliary Networks (SANs)

Images are dense 2D representations of the information captured by a camera, which makes convolutions a natural choice in most computer vision tasks [29]. Depth maps, however, are very sparse, often containing less than 1% valid pixels with useful information [20], thus making convolutions a sub-optimal choice as: (i) significant computational power is wasted on uninformative areas; (ii) spatial dependencies will include spurious information from these uninformative areas; and (iii) shared filters will still average loss gradients from the entire input depth map.

To avoid these shortcomings, we propose the use of sparse convolutions to process input depth maps, while RGB images are still processed using standard convolutions. More specifically, we use Minkowski convolutions [4], a highly efficient generalized sparse convolution recently introduced to address high-dimensional problems. In this work we focus on the 2D application of Minkowski convolutions (image processing), and leave potential higher-dimension applications (i.e., multi-view [16]

or spatio-temporal reasoning [4]) to future work. Within this framework, a sparse tensor S is written as a *coordinate matrix* C and a *feature matrix* F :

$$C = \begin{bmatrix} u_1 & v_1 & s_1 \\ \vdots & \vdots & \vdots \\ u_N & v_N & s_N \end{bmatrix}, \quad F = \begin{bmatrix} \mathbf{f}_1 \\ \vdots \\ \mathbf{f}_N \end{bmatrix}, \quad (5)$$

where $\{u_n, v_n\}$ are pixel coordinates, s_n is the sample index in the batch, and $\mathbf{f}_n \in \mathbb{R}^Q$ is the corresponding feature vector. For simplicity and without loss of generality, we assume a batch size of 1 and disregard the batch index. An input $W \times H \times 1$ depth map \tilde{D} is *sparsified* by gathering its valid pixels (i.e., with positive values) as coordinates and depth values as features, such that:

$$\tilde{S} = \left\{ \left\{ (u, v), \tilde{D}(u, v) \right\} \forall u, v \in \tilde{D} \mid \tilde{D}(u, v) > 0 \right\} \quad (6)$$

Similarly, a sparse tensor $\tilde{S} = \{\tilde{C}, \tilde{F}\}$ can be *densified* by scattering its pixel coordinates and feature values into a dense $W \times H \times Q$ matrix \tilde{P} , such that:

$$\tilde{P}(u_n, v_n) = \begin{cases} \mathbf{f}_n, & \text{if } \{u_n, v_n\} \in \tilde{C}. \\ \mathbf{0}, & \text{otherwise.} \end{cases} \quad (7)$$

Once the input depth map is sparsified, its information is encoded through a series of novel Sparse Residual Blocks (SRB), detailed in Fig. 2b. Each SRB is composed of three parallel branches that process the same input after an initial max pooling stage, each with a different number of sparse

convolutional blocks. The output of these branches is added and serves as input to the next SRB at a lower spatial resolution. Note that this entire chain of operations is sparse, and thus can be performed efficiently given the usually high sparsity of projected [7] or sampled [40] depth maps. After each block, a densification layer (Eq. 7) is used in parallel to generate a dense representation of these sparse features, that are then injected into the skip connections of the RGB module as detailed in the next section.

3.3. Proposed Architecture

Our proposed architecture for the joint learning of monocular depth prediction and completion is depicted in Fig. 2. It is composed of two modules, one for the processing of dense images (RGB) and one for the processing of sparse depth maps (SAN). The dense RGB module can be any encoder-decoder depth prediction network that uses skip connections [9, 15, 18, 30]. In our work we consider two baseline state-of-the-art network architectures: *PackNet* [18, 20] and *BTS* [30]. The sparse depth module uses our novel Sparse Residual Blocks described in Section 3.2 to encode sparsified depth maps used as input in conjunction with RGB images. Following the notation introduced in Section 3.1, the RGB module is defined by the parameters θ_I and the depth module is defined by θ_D .

If a single image I is used, only the RGB module is activated and the output will be a *predicted* depth map $f_P(I; \theta_I) = \hat{D}_P$. Alternatively, if sparse depth measurements \hat{D} are also provided, they serve as input to the SAN module, where they are encoded through a series of SRBs (Fig. 2b) to produce sparse depth features at increasingly lower resolutions. These resolutions are designed to match those of the RGB encoder, in such a way that the sparse depth features can be injected into the dense RGB features by simply adding the two feature maps, after densification. Because the network utilizes this sparse depth information in addition to the dense RGB image, its output will be a *completed* depth map $f_C(I, \hat{D}; \theta_I, \theta_D) = \hat{D}_C$.

Empirically, we have determined that injecting this information at the skip connection level is optimal to ensure that both tasks can still be performed by the same network without degradation. In this configuration, the RGB encoder only processes image features, while the RGB decoder processes features from the RGB encoder augmented with the sparse features from the depth encoder. To further condition the skip connections and enable the switching between tasks, we also introduce learnable parameters \mathbf{w} and \mathbf{b} as part of the SAN module. Assuming K_i as the feature map from the RGB encoder used as skip connection at scale resolution i , the augmented skip connection after introducing sparse depth information from \hat{P}_i is defined as:

$$\tilde{K}_i = w_i \times K_i + b_i + \hat{P}_i \quad (8)$$

Note that if no sparse depth information is available these parameters are not used. This enables the skip connections to be conditioned on the task being performed, and can better adapt to the introduction of additional information from the SAN module, minimizing gradient interference. A detailed study to determine the contribution of each component of our proposed architecture can be found in Tab. 3.

4. Experimental Protocol

4.1. Implementation Details

Our models¹ were implemented using Pytorch [36] and trained across eight V100 GPUs, with batch size $b = 4$ per GPU. We use the AdamW optimizer [32], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, starting learning rate $lr = 10^{-4}$ and weight decay $wd = 10^{-2}$. Our training schedule includes 30 epochs where only the depth prediction network is trained, followed by 20 epochs where the depth prediction encoder is frozen and only the depth completion encoder and shared decoder are trained. As training proceeds, the learning rate is decayed by a factor of 2 after every 20 epochs.

As the baseline depth prediction networks we considered *BTS* [30] and *PackNet* [18], using their official Pytorch implementations. With *BTS* we evaluate our architecture’s ability to improve upon the current state of the art in monocular depth prediction; and with *PackNet* we investigate whether a more complex architecture is better suited to simultaneously learn both tasks. Please note that our proposed Sparse Auxiliary Networks (SANs) can be equally applied to any other architecture, to benefit from potential improvements in speed, memory usage and performance.

4.2. Datasets

KITTI. We use the KITTI benchmark [11] and train on the *Eigen* split, composed of 23,488 training, 888 validation and 697 testing images (from which only 652 contain accumulated ground-truth depth maps [46]). Additionally, we present results on the KITTI public leaderboard, which consists of and 500 and 1,000 frames respectively for testing depth prediction and completion methods. Following standard procedure [30], at training time a random crop of 352×704 was used, with the addition of random horizontal flipping and color jittering.

DDAD. The Dense Depth for Automated Driving (DDAD) [18] is an urban driving dataset containing multiple synchronized cameras and depth ranges of up to 250 meters. It has a total of 12,560 training samples, from which we selected cameras 1/5/6/9 for a total of 50,600 images and ground-truth depth maps. The validation set contains 3,950 samples (15,800 images) and ground-truth depth maps. Following standard procedure [18], input images were down-sampled to a 640×384 resolution, and for evaluation we

¹Code available at: <https://github.com/TRI-ML/packnet-sfm>

Method	Input	Lower is better ↓					Higher is better ↑		
		Abs.Rel	Sqr.Rel	RMSE	RMSE _{log}	SILog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Kuznetsov et al. [28]	RGB	0.113	0.741	4.621	0.189	—	0.862	0.960	0.986
Gan et al. [10]	RGB	0.098	0.666	3.933	0.173	—	0.890	0.964	0.985
Guizilini et al. [20]	RGB	0.078	0.378	3.330	0.121	—	0.927	—	—
Fu et al. [9]	RGB	0.072	0.307	<u>2.727</u>	0.120	—	0.932	0.984	0.994
Yin et al. [53]	RGB	0.072	—	3.258	0.117	—	0.938	0.990	0.998
Lee et al. [30]	RGB	<u>0.059</u>	<u>0.245</u>	2.756	<u>0.096</u>	—	<u>0.956</u>	<u>0.993</u>	<u>0.998</u>
BTS-SAN	RGB	0.057	0.229	2.704	0.092	8.926	0.961	0.994	0.999
	RGB+D	0.021	0.038	1.094	0.037	3.749	0.996	0.999	1.000
PackNet-SAN	RGB	0.052	0.175	2.233	0.083	7.618	0.970	0.996	0.999
	RGB+D	0.015	0.028	0.909	0.032	3.149	0.997	0.999	1.000
<i>Improvement</i>	RGB	<i>11.9%</i>	<i>28.5%</i>	<i>18.9%</i>	<i>13.5%</i>	—	<i>1.4%</i>	<i>0.0%</i>	<i>0.0%</i>

Table 1: **Depth estimation results on the KITTI dataset**, for the *Eigen* test split [7] and distances up to 80m. The *Improvements* row indicate the percentual improvement between our best model (PackNet-SAN) and the current state of the art (BTS, by Lee et al. [30], underlined).

considered distances up to 200m without any cropping. A single model was trained using all four cameras, and evaluated individually on each one.

NYUv2. To evaluate our proposed methodology on other domains, we also provide results on the NYUv2 dataset [40]. It consists of RGB+D data collected from 464 scenes, with 249 used for training and 215 for testing. We follow [30] and sample frames evenly from the training sequences, generating roughly 36k training RGB+D images. For depth prediction we train on images of size 640×480 , while for depth completion we first downsample the original frames by half and center-crop to 304×228 , so as to be consistent with the protocol followed by related methods [33]. Additionally, for depth completion we use input depth maps with 200 or 500 valid points respectively, randomly sampled from the original depth images, following the standard training protocol on this dataset [33]. We upsample each test prediction to the original test image resolution and evaluate on a center crop following related work [1, 9, 30], using the official test split of 654 frames.

KITTI3D. To further analyze the accuracy of the depth maps predicted by our proposed SAN architecture, we also evaluated their performance in the downstream task of monocular 3D object detection as *pseudo-LiDAR* point-clouds. Specifically, we use the KITTI3D dataset [12], composed of 3,712 training and 3,712 validation images.

Pretraining. Following related work [30, 33, 47], we found pre-training to improve network performance. For our KITTI experiments we pre-train on a larger split of DDAD, while for the NYU experiments we pre-train on the Scannet dataset [5] by sampling approximately 250k RGB+D frames without any additional cropping or filtering. We ablate the effect of pretraining in Tab. 3.

	Method	SILog	SqRel	AbsRel	iRMSE
Prediction	SGDepth [26]	15.30	5.00%	13.29%	15.80
	SDNet [35]	14.68	3.90%	12.31%	15.96
	VGG26-UNet [21]	13.41	2.86%	10.60%	15.06
	PAP [56]	13.08	2.72%	10.27%	13.95
	VNL [50]	12.65	2.46%	10.15%	13.02
	SORD [6]	12.39	2.49%	10.10%	13.48
	RefinedMPL [48]	11.80	2.31%	10.09%	13.39
	DORN [9]	11.77	2.23%	8.78%	12.98
	BTS [30]	11.67	2.21%	9.04%	12.23
	PackNet-SAN	11.54	2.35%	9.12%	12.38
	Method	RMSE	iRMSE	MAE	iMAE
Completion	DCDC [24]	1109.04	2.95	234.01	1.07
	CSPN [3]	1019.64	2.93	279.46	1.15
	Conf-Net [22]	962.28	3.10	257.54	1.09
	Sparse-to-Dense [33]	954.36	3.21	288.64	1.35
	DFineNet [55]	943.89	1.39	304.17	1.39
	CrossGuidance [31]	807.42	2.73	253.98	1.33
	FusionNet [47]	772.87	2.19	215.02	0.93
	GuideNet [42]	736.24	2.25	218.83	0.99
	PackNet-SAN	914.35	2.78	298.04	1.36

Table 2: **Depth estimation results on the official KITTI testset benchmark** relative to other published methods, for both prediction and completion tasks (bold metrics are used for leaderboard scoring). Note that the same model was used in both submissions, only modifying the input information (RGB for prediction and RGB+D for completion).

5. Experimental Results

5.1. Depth Prediction and Completion

KITTI. In Tab. 1 we present quantitative results for the tasks of depth prediction and completion, considering the *Eigen* test split. We note that *BTS-SAN*, i.e. the BTS archi-

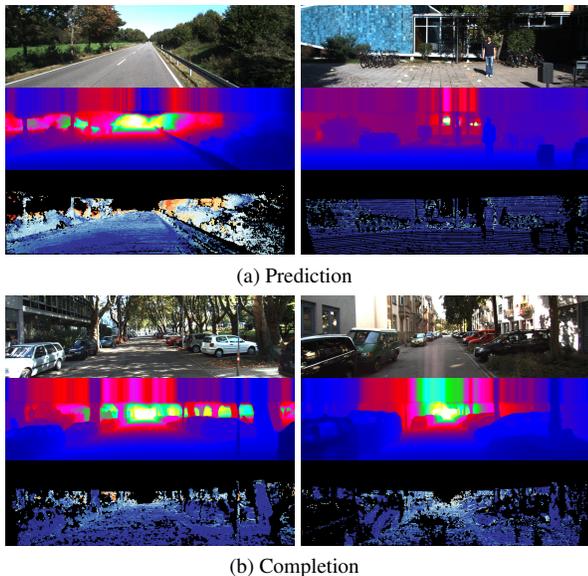


Figure 3: **Qualitative depth prediction and completion results on the KITTI benchmark**, using PackNet-SAN.

texture [30] with our proposed SAN module improves over the baseline numbers for the task of depth prediction, while at the same time enabling depth completion if sparse depth maps are also provided as additional input. These results are further improved by using PackNet [18] as the underlying depth prediction network, establishing a new state of the art for this task by a significant margin. We also evaluated our proposed *PackNet-SAN* architecture on the official KITTI test set benchmark, submitting results from the same model to both depth prediction and completion leaderboards (see Tab. 2). Despite operating in this challenging setting, at the time of publication our method ranked first amongst all published methods for the task of depth prediction for the *SILog* metric (used to determine ranking), while at the same time showing good depth completion performance. We show qualitative results obtained from the KITTI leaderboard in Fig. 3.

DDAD. In Tab. 5 we show results on the DDAD dataset obtained using our baseline depth prediction network (PackNet) and its extension using our proposed SAN architecture, to enable the joint-task learning of depth prediction and completion. From these results we note that the introduction of joint-task learning boosted depth prediction results by a significant margin, similarly to what was observed in the KITTI experiments (for qualitative results, see Fig. 4). Additionally, if sparse depth maps are available as input they can also be used to generate depth completion results, further improving performance. We note that the RGB+D experiments on DDAD were performed with a sparsity level of 20% for the input depth maps - we provide a detailed analysis of how the sparsity level affects performance in the ablative section (see Fig. 6).

Method	Input	Abs.Rel↓	RMSE↓	SILog↓	$\delta < 1.25\uparrow$
SRB x1	RGB	0.057	2.483	8.064	0.966
	RGB+D	0.019	0.994	3.343	0.997
SRB x2	RGB	0.055	2.328	7.862	0.966
	RGB+D	0.017	0.949	3.287	0.997
Unfreeze Pred. Encoder	RGB	0.055	2.306	7.978	0.967
	RGB+D	0.021	0.965	3.333	0.996
Freeze Pred. Decoder	RGB	0.054	2.318	7.901	0.968
	RGB+D	0.024	1.070	3.805	0.995
W/o W_i and B_i parameters	RGB	0.056	2.374	8.324	0.962
	RGB+D	0.019	0.958	3.395	0.995
Train from Scratch	RGB	0.062	2.888	9.579	0.955
	RGB+D	0.019	1.049	3.631	0.996
Prediction	RGB	0.054	2.476	8.081	0.966
Completion	RGB+D	0.015	0.878	3.238	0.997
PackNet-SAN	RGB	0.052	2.233	7.618	0.970
	RGB+D	0.015	0.909	3.149	0.997

Table 3: **Ablation analysis on the KITTI dataset**, considering the *Eigen* test split [7] and *PackNet* [18] as the depth prediction network. *SRB xX* uses Sparse Residual Blocks with fewer branches; *Unfreeze Pred. Encoder* also updates the prediction encoder during the second stage of training; *Freeze Pred. Decoder* also freezes the prediction decoder during the second stage of training; *w/o W_i and B_i* removes the shared parameters for each skip connection; *Train from scratch* does not use a pre-trained model; and *Prediction* and *Completion* are trained only for that particular task.

NYUv2. Our NYUv2 results are summarized in Tab. 6, observing the same trend as on the other datasets. The proposed architecture *PackNet-SAN* improves over the baseline PackNet [18], achieving new state-of-the-art performance for depth prediction on this dataset. When using RGB+D data at inference time, our method is competitive with state-of-the-art methods, achieving similar numbers on most metrics. We show qualitative results on NYUv2 in Fig. 5.

5.2. Monocular 3D Object Detection

To further analyze the accuracy of the depth maps predicted by our proposed SAN architecture, we evaluated their performance in the downstream task of monocular 3D

Method	AP3D@easy	AP3D@medium	AP3D@hard
DORN [9]	34.8/35.1	22.0/22.0	19.5/19.6
PackNet-SAN	35.5/35.7	22.6/22.8	19.9/20.1

Table 4: **3D object detection results on the validation set of KITTI3D for the Car category**, using PatchNet [34] and different monocular pointclouds (no input sparse depth), for the *validation* split. The same detection architecture and learning hyperparameters were used in both cases.

	Input	Camera	Lower is better ↓					Higher is better ↑		
			Abs.Rel	Sqr.Rel	RMSE	RMSE _{log}	SILog	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
PackNet	RGB	01	0.088	1.760	11.331	0.195	18.499	0.899	0.960	0.981
		05	0.130	2.025	10.472	0.268	25.273	0.832	0.927	0.960
		06	0.151	2.485	10.680	0.307	28.007	0.791	0.904	0.944
		09	0.132	2.362	12.497	0.261	24.551	0.821	0.925	0.962
PackNet-SAN	RGB	01	0.083	1.575	10.693	0.185	17.767	0.911	0.967	0.987
		05	0.127	1.863	10.210	0.263	24.966	0.841	0.931	0.973
		06	0.145	2.307	10.493	0.298	27.491	0.804	0.911	0.968
		09	0.119	1.979	12.010	0.256	24.295	0.844	0.936	0.978
		Avg. Improv.	5.45%	10.47%	3.44%	2.96%	2.01%	1.71%	0.78%	1.54%
PackNet-SAN	RGB+D	01	0.052	0.933	8.683	0.153	14.920	0.955	0.978	0.987
		05	0.072	1.097	7.950	0.207	20.375	0.928	0.958	0.973
		06	0.081	1.255	7.994	0.232	22.675	0.922	0.955	0.969
		09	0.067	1.131	9.052	0.189	18.481	0.934	0.966	0.979

Table 5: **Depth estimation results on DDAD** using PackNet-SAN, with the same model trained on four cameras considering distances up to 200m. For the RGB+D experiments, a sparsity level of 20% was used for input depth maps (see Fig. 6).

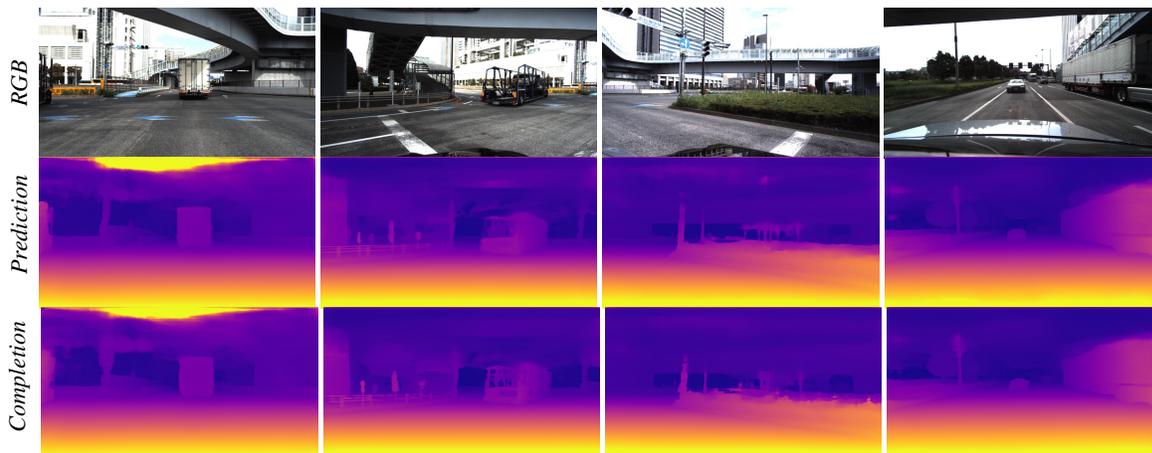


Figure 4: **Qualitative depth results on DDAD**, using PackNet-SAN. A single sample is shown, each column corresponding to an individual camera. The same model was trained on all four cameras both for the task of depth prediction (middle row) and completion (bottom row), as shown in Tab. 5.

object detection, using the recently proposed PatchNet architecture [34]. The depth maps predicted by *PackNet-SAN* were projected into 3D as *pseudo-LiDAR* pointclouds using ground-truth camera intrinsics, and then used as input to PatchNet without any other modifications. In Tab. 4 we present results on the KITTI3D dataset and show that by operating on our pointclouds we increase object detection performance in all difficulty thresholds relative to the previous state of the art when using the DORN [9] depth estimator. Note that, for a fair comparison, the two depth networks (DORN and PackNet-SAN) were trained using the same *Eigen* split of KITTI.

5.3. Ablative Analysis

In Tab. 3 we perform a comprehensive ablation study showing the effects of each design choice of our proposed

architecture, and how they contribute to these improvements. In particular we show that the joint learning of both tasks actually improves depth prediction performance relative to single task learning, without degrading depth completion performance. We also demonstrate that increasing the complexity of the sparse encoder (i.e. introducing more sparse residual blocks) benefits both tasks, since it facilitates the decoupling of RGB and depth features without overloading the shared decoder. We also experimented with different schedules for parameter freezing, and determined that freezing the dense encoder after the initial depth prediction learning stage leads to optimal results.

Additionally, in Fig. 6 we analyze the impact of sparsity in the input depth maps on the *DDAD* dataset. Specifically, we sparsify the input depth maps by randomly sampling a percentage of valid input depth pixels (depth maps used

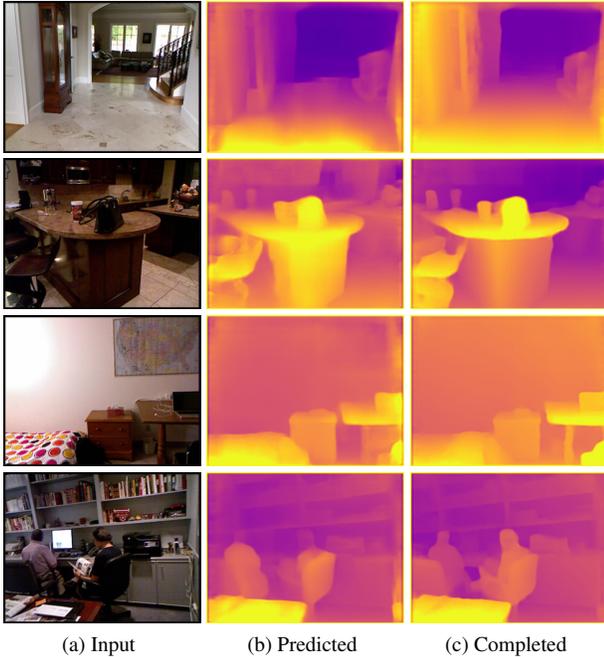


Figure 5: **Qualitative depth results on NYUv2**, using PackNet-SAN. Our joint-task learning methodology enables the generation of state-of-the-art predicted depth maps, that can be further improved by using sparse depth maps as additional input without changing the architecture.

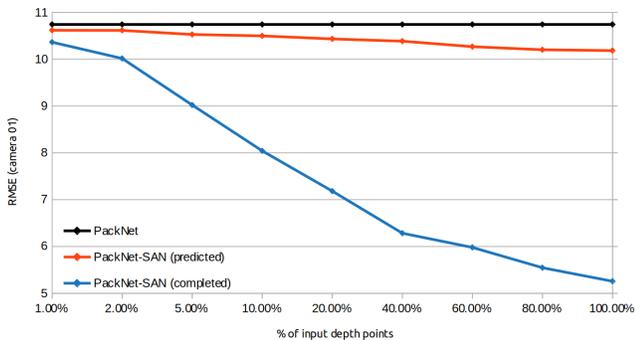


Figure 6: **Sparsity analysis on the DDAD dataset**, using PackNet-SAN. We show depth prediction (red) and completion (blue) results with different levels of sparsity, compared to the baseline prediction-only network (black).

for supervision and evaluation were not modified). As expected, performance increases with the percentage of available input depth points, showing that our proposed SAN architecture is able to leverage different levels of sparsity to consistently improve results. Interestingly, we observe a similar trend for depth prediction results as well, as further evidence that joint-task learning of depth prediction and completion is able to further improve results even when only RGB images are utilized at test time.

Depth Prediction					
Method	AbsRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Qi et al. [38]	0.128	0.569	0.834	0.960	0.990
Alhashim et al. [1]	0.123	0.465	0.846	0.974	0.994
Fu et al. [9]	0.115	0.509	0.828	0.965	0.992
Yin et al. [53]	0.108	0.416	0.875	0.976	0.994
Lee et al. [30]	0.110	0.392	0.885	0.978	0.994
PackNet [18]	0.110	0.397	0.886	0.979	0.995
PackNet-SAN	0.106	0.393	0.892	0.979	0.995
Depth Completion - 200 samples					
Ma et al. [33]†	0.044	0.230	0.971	0.994	0.998
NConv-CNN [8]†	0.027	0.173	0.982	0.996	0.999
Tang et al. [42]	0.024	0.142	0.988	0.998	1.000
PackNet-SAN	0.027	0.155	0.989	0.998	0.999
Depth Completion - 500 samples					
Ma et al. [33]	0.043	0.204	0.978	0.996	0.999
DeepLidar [39]	0.022	0.115	0.993	0.999	1.000
EncDec-Net[EF] [8]	0.017	0.123	0.991	0.998	1.000
CSPN [3]	0.016	0.117	0.992	0.999	1.000
Tang et al. [42]	0.015	0.101	0.995	0.999	1.000
PackNet-SAN	0.019	0.120	0.994	0.999	1.000

Table 6: **Depth estimation results on the test split of the NYUv2 dataset**, relative to other published methods, for both depth prediction and completion tasks. Note that the same model was used in both submissions, the only modification being the input information (RGB for prediction and RGB+D for completion). † - results from [42].

6. Conclusion

This paper describes a novel methodology for monocular depth estimation that combines the tasks of depth prediction and completion into a single architecture. We propose a mid-level fusion approach for the joint learning of both tasks, using a standard depth prediction network with the addition of a sparse encoder to process input depth maps. The sparse depth features are added to the skip connections of the image encoder at each layer, before they are fed into a shared dense decoder. The resulting architecture can be used to perform both tasks without any further training, simply by modifying the input information between RGB and RGB+D or controlling the sparsity level of the input depth maps. Through an extensive analysis on different benchmarks, we demonstrate that our proposed unified SAN architecture achieves a new state of the art in monocular depth prediction by a significant margin. As future work, we will explore multi-frame extensions (e.g., stereo pairs or temporal context), as well as developing ways to further improve depth completion performance in the SAN setting.

References

- [1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. [5](#), [8](#)
- [2] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *arXiv preprint arXiv:1908.10553*, 2019. [1](#), [2](#)
- [3] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119, 2018. [5](#), [8](#)
- [4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. [2](#), [3](#)
- [5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. [5](#)
- [6] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4738–4747, 2019. [5](#)
- [7] 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. [1](#), [2](#), [4](#), [5](#), [6](#)
- [8] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2019. [2](#), [8](#)
- [9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [10] Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *ECCV*, 2018. [1](#), [5](#)
- [11] 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. [2](#), [4](#)
- [12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. [2](#), [5](#)
- [13] Xavier Glorot, Antoine Bordes, and Y. Bengio. Deep sparse rectifier neural networks. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 315–323, 2011. [3](#)
- [14] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. [1](#), [2](#)
- [15] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *ICCV*, 2019. [1](#), [4](#)
- [16] Zan Gojcic, Caifa Zhou, Jan D Wegner, Leonidas J Guibas, and Tolga Birdal. Learning multiview 3d point cloud registration. In *International conference on computer vision and pattern recognition (CVPR)*, 2020. [3](#)
- [17] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8977–8986, 2019. [1](#)
- [18] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Rantos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [1](#), [2](#), [3](#), [4](#), [6](#), [8](#)
- [19] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. [2](#)
- [20] Vitor Guizilini, Jie Li, Rares Ambrus, Sudeep Pillai, and Adrien Gaidon. Robust semi-supervised monocular depth estimation with reprojected distances. In *Conference on Robot Learning (CoRL)*, October 2019. [1](#), [2](#), [3](#), [4](#), [5](#)
- [21] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. [2](#), [5](#)
- [22] Hamid Hekmatian, Jingfu Jin, and Samir Al-Stouhi. Confnet: Toward high-confidence dense 3d point-cloud with error-map prediction. *arXiv preprint arXiv:1907.10148*, 2019. [5](#)
- [23] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. *arXiv preprint arXiv:2003.06409*, 2020. [1](#)
- [24] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris. Depth coefficients for depth completion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12438–12447. IEEE, 2019. [1](#), [2](#), [5](#)
- [25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, page 448–456, 2015. [3](#)
- [26] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. *arXiv preprint arXiv:2007.06936*, 2020. [2](#), [5](#)
- [27] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: Learning sfm from sfm. In *European Conference on Computer Vision*, pages 713–728. Springer, 2018. [2](#)

- [28] Yevhen Kuznetsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017. 5
- [29] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 3
- [30] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 2, 3, 4, 5, 6, 8
- [31] Sihaeng Lee, Janghyeon Lee, Doyeon Kim, and Junmo Kim. Deep architecture with cross guidance between single image and sparse lidar data for depth completion. *IEEE Access*, 8:79801–79810, 2020. 2, 5
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4
- [33] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018. 1, 2, 5, 8
- [34] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6, 7
- [35] Matthias Ochs, Adrian Kretz, and Rudolf Mester. Sdnet: Semantically guided depth estimation network. In *German Conference on Pattern Recognition*, pages 288–302. Springer, 2019. 2, 5
- [36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 4
- [37] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020. 2
- [38] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 2, 8
- [39] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2019. 1, 8
- [40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 2, 4, 5
- [41] Jiexiong Tang, Rares Ambrus, Vitor Guizilini, Sudeep Pillai, Hanme Kim, and Adrien Gaidon. Self-supervised 3d keypoint learning for ego-motion estimation. *arXiv preprint arXiv:1912.03426*, 2019. 2
- [42] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *arXiv preprint arXiv:1908.01238*, 2019. 2, 5, 8
- [43] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998. 2
- [44] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4665, 2020. 2
- [45] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 2
- [46] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. *3DV*, 2017. 4
- [47] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019. 2, 5
- [48] Jean Marie Uwabeza Vianney, Shubhra Aich, and Bingbing Liu. Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving. *arXiv preprint arXiv:1911.09712*, 2019. 5
- [49] Jamie Watson, Michael Firman, Aron Monzspart, and Gabriel J Brostow. Footprints and free space from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2020. 1
- [50] Yin Wei, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. *arXiv preprint arXiv:1907.12209*, 2019. 2, 5
- [51] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1281–1292, 2020. 1, 2
- [52] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. *arXiv preprint arXiv:1807.02570*, 2018. 2
- [53] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 8
- [54] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018. 2
- [55] Yilun Zhang, Ty Nguyen, Ian D Miller, Shreyas S Shivakumar, Steven Chen, Camillo J Taylor, and Vijay Kumar.

- Dfinenet: Ego-motion estimation and depth refinement from sparse, noisy depth input with rgb guidance. *arXiv preprint arXiv:1903.06397*, 2019. [2](#), [5](#)
- [56] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019. [2](#), [5](#)
- [57] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020. [2](#)
- [58] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *ECCV*, pages 6871–6880, 10 2019. [1](#)
- [59] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. [2](#)
- [60] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. [1](#)