

Dense Depth Posterior (DDP) from Single Image and Sparse Range

Yanchao Yang, Alex Wong, Stefano Soatto
 UCLA Vision Lab
 University of California, Los Angeles, CA 90095
 {yanchao.yang, alexw, soatto}@cs.ucla.edu

Abstract

We present a deep learning system to infer the posterior distribution of a dense depth map associated with an image, by exploiting sparse range measurements, for instance from a lidar. While the lidar may provide a depth value for a small percentage of the pixels, we exploit regularities reflected in the training set to complete the map so as to have a probability over depth for each pixel in the image. We exploit a Conditional Prior Network, that allows associating a probability to each depth value given an image, and combine it with a likelihood term that uses the sparse measurements. Optionally we can also exploit the availability of stereo during training, but in any case only require a single image and a sparse point cloud at run-time. We test our approach on both unsupervised and supervised depth completion using the KITTI benchmark, and improve the state-of-the-art in both. Code is available at: <https://github.com/YanchaoYang/Dense-Depth-Posterior>

1. Introduction

There are many dense depth maps that are compatible with a given image and a sparse point cloud. Any point-estimate, therefore, depends critically on the prior assumptions made. Ideally, one would compute the entire posterior distribution of depth maps, rather than a point-estimate. The posterior affords to reason about confidence, integrating evidence over time, and in general, is a (Bayesian) sufficient representation that accounts for all the information in the data.

Motivating application. In autonomous navigation, a sparse point cloud from lidar may be insufficient to make planning decisions: Is the surface of the road in Fig. 1 (middle, better viewed when enlarged) littered with pot-holes, or is it a smooth surface? Points that are nearby in image topology, projecting onto adjacent pixels, may be arbitrarily far in the scene. For instance, pixels that straddle an occluding boundary correspond to large depth gaps in the scene. While the lidar may not measure every pixel, if we

know it projects onto a tree, trees tend to stand out from the ground, which informs the topology of the scene. On the other hand, pixels that straddle illumination boundaries, like shadows cast by trees, seldom correspond to large depth discontinuities.

Depth completion is the process of assigning a depth value to each pixel. While there are several deep learning-based methods to do so, we wish to have the entire posterior estimate over depths. Sparse range measurements serve to ground the posterior estimate in a metric space. This could then be used by a decision and control engine downstream.

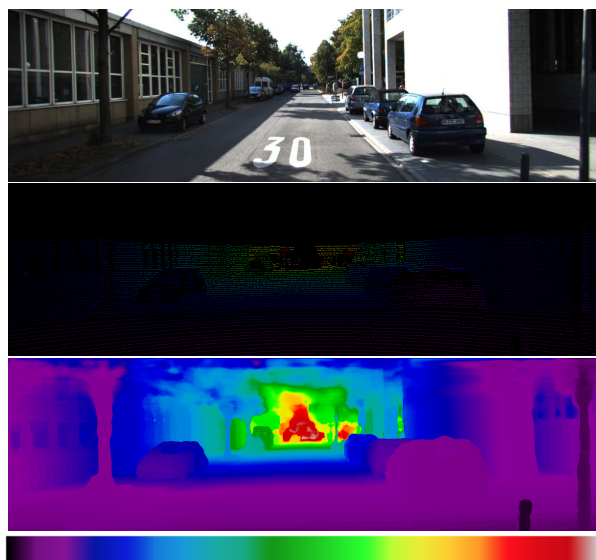


Figure 1. An image (top) is insufficient to determine the geometry of the scene; a point cloud alone (middle) is similarly ambiguous. Lidar returns are shown as colored points, but black regions are uninformative: Are the black regions holes in the road surface, or due to radiometric absorption? Combining a single image, the lidar point cloud, and previously seen scenes allows inferring a dense depth map (bottom) with high confidence. Color bar from left to right: zero to infinity.

Side information. If the dense depth map is obtained by processing the given image and sparse point cloud alone, the quality of the resulting decision or control action could

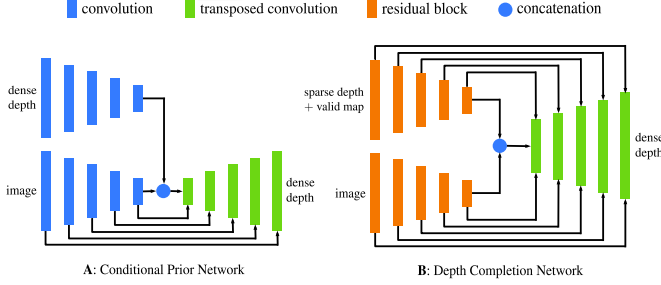


Figure 2. (A): the architecture of the Conditional Prior Network (CPN) to learn the conditional of the dense depth given a single image. (B): Our proposed Depth Completion Network (DCN) for learning the mapping from a sparse depth map and an image to a dense depth map. Connections within each encoder/decoder block are omitted for simplicity.

be no better than if the raw data was fed downstream (Data Processing Inequality). However, if depth completion can exploit a prior or aggregate experience from previously seen images and corresponding dense depth maps, then it is possible for the resulting dense depth map to improve the quality of the decision or action, assuming that the training set is representative. To analyze a depth completion algorithm, it is important to understand what prior assumptions, hypotheses or side information is being exploited.

Goal. We seek methods to *estimate the geometry and topology of the scene given an image, a sparse depth map, and a body of training data consisting of images and the associated dense depth maps*. Our assumption is that the distribution of seen images and corresponding depth maps is representative of the present data (image and sparse point cloud) once restricted to a sparse domain.

Our method yields the full posterior over depth maps, which is much more powerful than any point estimate. For instance, it allows reasoning about confidence intervals. We elect the simplest point estimate possible, which is the maximum, to evaluate the accuracy of the posterior. It should be noted, however, that when there are multiple hypotheses with similar posterior, the point estimate could jump from one mode to another, and yet the posterior being an accurate representation of the unknown variable. More sophisticated point estimators, for instance, taking into account memory, or spatial distribution, non-maximum suppression, etc. could be considered, but here we limit ourselves to the simplest one.

Key idea. While an image alone is insufficient to determine a depth map, certain depth maps are more probable than others given the image and a previously seen dataset. The key to our approach is a conditional prior model $P(d|I, \mathcal{D})$ that scores the compatibility of each dense depth map d with the given image I based on the previously observed dataset \mathcal{D} . This is computed using a Conditional Prior Network (CPN) [36] in conjunction with a model of

the likelihood of the observed sparse point cloud z under the hypothesized depth map d , to yield the posterior probability and, from it, a maximum a-posteriori (MAP) estimate of the depth map for benchmark evaluation:

$$\hat{d} = \arg \max_d P(d|I, z) \propto P(z|d)P_{\mathcal{D}}(d|I). \quad (1)$$

Let $D \subset \mathbb{R}^2$ be the image domain, sampled on a regular lattice of dimension $N \times M$, $I : D \rightarrow \mathbb{R}^3$ is a color image, with the range quantized to a finite set of colors, $d : D \rightarrow \mathbb{R}_+$ is the dense depth map defined on the lattice D , which we represent with an abuse of notation as a vector of dimension MN : $d \in \mathbb{R}_+^{NM}$. $\Omega \subset D$ is a sparse subset of the image domain, with cardinality $K = |\Omega|$, where the function d takes values $d(\Omega) = z \in \mathbb{R}_+^K$. Finally, $\mathcal{D} = \{d_j, I_j\}_{j=1}^n$ is a dataset of images I_j and their corresponding dense depth maps $d_j \in \mathbb{R}_+^{NM}$. Since we do not treat \mathcal{D} as a random variable but a given set of data, we write it as a subscript. In some cases, we may have additional data available during training, for instance stereo imagery, in which case we include it in the dataset, and discuss in detail how to exploit it in Sect. 3.3.

Results. We train a deep neural network model to produce an estimate of the posterior distribution of dense depth maps given an image and a sparse point cloud (sparse range map), that leverages a Conditional Prior Network to restrict the hypothesis space, weighted by a classical likelihood term. We use a simple maximum a-posteriori (MAP) estimate to evaluate our approach on benchmark datasets, including the KITTI-unsupervised, where the dense depth map is predicted given an image and a point cloud with 5% pixel coverage, and the KITTI-supervised, where a point cloud with 30% coverage is given for training. We achieve top performance in both. We also validate on additional data in the Supplementary Materials [37].

2. Related Work

Semi-Dense Depth Completion. Structured light sensors typically provide dense depth measurements with about 20% missing values; At this density, the problem is akin to inpainting [2, 20, 27] that use morphological operations [18, 24]. The regime we are interested in involves far sparser point clouds (> 90% missing values).

Supervised Depth Completion. Given a single RGB image and its associated sparse depth measurements along with dense ground truth, learning-based methods [7, 15, 25, 29, 38] minimize the corresponding loss between prediction and ground truth depth. [29] trains a deep network to regress depth using a sparse convolutional layer that discounts the invalid depth measurements in the input while [15] proposes a sparsity-invariant upsampling layer, sparsity-invariant summation, and joint sparsity-invariant concatenation and convolution. [7] treats the binary validity map as a confidence map and adapts normalized convo-

lution for confidence propagation through layers. [5] implements an approximation of morphological operators using the contra-harmonic mean (CHM) filter [23] and incorporates it as a layer in a U-Net architecture for depth completion. [4] proposes a deep recurrent auto-encoder to mimic the optimization procedure of compressive sensing for depth completion, where the dictionary is embedded in the neural network. [38] predicts surface normals and occlusion boundaries from the RGB image, which gives a coarse representation of the scene structure. The predicted surface normals and occlusion boundaries are incorporated as constraints in a global optimization framework guided by sparse depth.

Unsupervised Depth Completion. In this problem setting, dense ground truth depth is not available as supervision, so a strong prior is key. [21] proposes minimizing the photometric consistency loss among a sequence of images with a second-order smoothness prior based on a similar formulation in single image depth prediction [22, 30, 39]. Instead of having a separate pose network or using direct visual odometry methods, [21] uses Perspective-n-Point (PnP) [19] and Random Sample Consensus (RANSAC) [9] to obtain pose. We exploit recently introduced method to learn the conditional prior [36] to take into account scene semantics rather than using a local smoothness assumption.

Stereo as Supervision. Recent works in view synthesis [10, 33] and unsupervised single image depth prediction, [8, 12, 13, 32] propose using view synthesis to hallucinate a novel view image by reconstruction loss. In the case of stereo pairs, [12, 13, 32] propose training networks to predict the disparities of an input image by reconstructing the unseen right view of a stereo pair given the left. In addition to the photometric reconstruction loss, local smoothness is assumed; [13] additionally proposed edge-aware smoothness and left-right consistency. Although during inference, we assume only one image is given, at training time we may have stereo imagery available, which we exploit as in Sect. 3.3. In this work, we incorporate only the stereo photometric reconstruction term. Despite our network predicting depths and the network [12, 13, 32] predicting disparities, we are able to incorporate this training scheme seamlessly into our approach.

Exploiting Semantics and Contextual Cues. While methods [7, 15, 21, 25, 29, 38] learn a representation for the depth completion task through ground truth supervision, they do not have any explicit modeling of the semantics of the scene. Recently, [26] explored this direction by predicting object boundary and semantic labels through a deep network and using them to construct locally planar elements that serve as input to a global energy minimization for depth completion. [3] proposes to complete the depth by anisotropic diffusion with a recurrent convolution network, where the affinity matrix is computed locally from an im-

age. [16] also trains a U-Net for joint depth completion and semantic segmentation in the form of multitask learning in an effort to incorporate semantics in the learning process.

To address contextual cues and scene semantics, [36] introduces a Conditional Prior Network (CPN) in the context of optical flow, which serves as a learning scheme for inferring the distribution of optical flow vectors given a single image. We leverage this technique and formulate depth completion as a maximum a-posteriori problem by factorizing it into a likelihood term and a conditional prior term, making it possible to explicitly model the semantics induced regularity of a single image. Even though our method could be applied to sparse-to-dense interpolation for optical flow, where the sparse matches can be obtained using [35, 34], here we focus our test on depth completion task.

3. Method

In order to exploit a previously observed dataset \mathcal{D} , we use a Conditional Prior Network (CPN) [36] in our framework. Conditional Prior Networks infer the probability of an optical flow given a single image. During training, ground truth optical flow is encoded (upper branch in Fig. 2-A), concatenated with the encoder of an image (lower branch), and then decoded into a reconstruction of the input optical flow.

In our implementation, the upper branch encodes dense depth, concatenated with the encoding of the image, to produce a dense reconstruction of depth at the decoder, together with a normalized likelihood that can serve as a posterior score. We consider a CPN as a function that, given an image (lower branch input) maps any sample putative depth map (upper branch input) to a positive real number, which represents the conditional probability/prior of the input dense depth map given the image.

We denote the ensemble of parameters in the CPN as w^{CPN} ; with abuse of notation, we denote the decoded depth with $d' = w^{CPN}(d, I)$. When trained with a bottleneck imposed on the encoder (upper branch), the reconstruction error is proportional to the conditional distribution:

$$Q(d, I; w^{CPN}) = e^{-\|w^{CPN}(d, I) - d\|^\eta} \propto P_{\mathcal{D}}(d|I) \quad (2)$$

where η indicates the specific norm used for calculating Q . In Sect. 4.2 and Sect. 5, we show the training details of CPN, and also quantitatively show the effect of different choices of the norm η . In the following, we assume w^{CPN} is trained, and Q will be used as the conditional prior. For the proof that Q computed by CPN represents the conditional prior as in Eq. (2), please refer to [36].

In order to obtain a posterior estimate of depth, the CPN needs to be coupled with a likelihood term.

3.1. Supervised Single Image Depth Completion

Supervised learning of dense depth assumes the availability of ground truth dense depth maps. In the KITTI depth completion benchmark [29], these are generated by accumulating the neighboring sparse lidar measurements. Even though it is called ground truth, the density is only $\sim 30\%$ of the image domain, whereas the density of the unsupervised benchmark is $\sim 5\%$. The training loss in the supervised modality is just the prediction error:

$$L(w) = \sum_{j=1}^N \|\phi(z_j, I_j; w) - d_j\|^\gamma \quad (3)$$

where ϕ is the map from sparse depth z and image I to dense depth, realized by a deep neural network with parameters w , and $\gamma = 1$ fixed in the supervised training.

Our network structure for ϕ is detailed in Fig. 2-B, which has a symmetric two-branch structure, each encoding different types of input: one sparse depth, the other an image; skip connections are enabled for two branches. Note that our network structure is unique among all the top performing ones on the KITTI depth completion benchmark: We do not use specifically-designed layers for sparse inputs, such as sparsity invariant layers [15, 29]. Instead of early fusion of sparse depth and image, our depth defers fusion to decoding, which entails fewer learnable parameters, detailed in [37]. A related idea is proposed in [16]; instead of a more sophisticated NASNet block [40], we use the more common ResNet block [14]. Although simpler than competing methods, our network achieves state-of-the-art performance (Sect. 5).

3.2. Unsupervised Single Image Depth Completion

Supervised learning requires ground truth dense depth, which is hard to come by. Even the “ground truth” provided in the KITTI benchmark is only 30% dense and interpolated from even sparser maps. When only sparse independent measurements of depth are available, for instance from lidar, with less than 10% coverage (e.g. 5% for KITTI), we call depth completion *unsupervised* as the only input are sensory data, from images and a range measurement device, with no annotation or pre-processing of the data.

The key to our approach is the use of a CPN to score the compatibility of each dense depth map d with the given image I based on the previously observed data \mathcal{D} . In some cases, we may have additional sensory data available *during training*, for instance, a second image taken with a camera with a known relative pose, such as stereo. In this case, we include the reading from the second camera in the training set \mathcal{D} , as described in Sect. 3.3. When only a single image is given, the CPN Eq. (2) is combined with a model of the likelihood of the observed sparse point cloud z under the

hypothesized depth map d :

$$P(z|d) \propto e^{-\|z-d(\Omega)\|^\gamma} \quad (4)$$

which is simply a Gaussian around the hypothesized depth, restricted to the sparse subset Ω , when $\gamma = 2$. The overall loss is:

$$\begin{aligned} L^u(w) &= - \sum_{j=1}^N \log P(d_j | I_j, z_j, \mathcal{D}) \\ &= \sum_{j=1}^N \|z_j - d_j(\Omega)\|^\gamma + \alpha \sum_{j=1}^N \|w^{CPN}(d_j, I_j) - d_j\|^\eta \\ &= \sum_{j=1}^N \|z_j - \phi(z_j, I_j; w)(\Omega)\|^\gamma + \\ &\quad \alpha \sum_{j=1}^N \|w^{CPN}(\phi(z_j, I_j; w), I_j) - \phi(z_j, I_j; w)\|^\eta \end{aligned} \quad (5)$$

Note that γ, η control the actual norm used during training, as well as the modeling of the likelihood and conditional distribution. We experiment with these parameters in Sect. 5.1, and show our quantitative analysis there.

3.3. Disparity Supervision

Some datasets come with stereo imagery. We want to be able to exploit it, but without having to require its availability at inference time. We exploit the strong relation between depth and disparity. In addition to the sparse depth z and the image I , we are given a second image I' as part of a stereo pair, which is rectified (standard pre-processing), to first-order we assume that there exists a displacement $s = s(x), x \in D$ such that

$$I(x) \approx I'(x + s) \quad (6)$$

which is the intensity constancy constraint. We model, again simplistically, disparity s as $s = FB/d$, where F is the focal length and B is the baseline (distance between the optical centers) of the cameras. Hence, we can synthesize disparity s from the predicted dense depth d , thus to constrain the recovery of 3-d scene geometry. More specifically, we model the likelihood of seeing I' given I, d as:

$$P(I'|I, d) \propto e^{-\frac{\sum_x \|I(x) - I'(x + s(d(x)))\|^2}{\delta^2}} \quad (7)$$

However, the validity of the intensity constancy assumption is affected by complex phenomena such as translucency, transparency, inter-reflection, etc. In order to mitigate the error in the assumption, we could also employ a perceptual metric of structural similarity (SSIM) [31]. SSIM scores corresponding 3×3 patches $p(x), p'(x) \in$

Method	iRMSE	iMAE	RMSE	MAE	Rank
Dimitrievski [6]	3.84	1.57	1045.45	310.49	13.0
Cheng [3]	2.93	1.15	1019.64	279.46	7.5
Huang [15]	2.73	1.13	841.78	253.47	6.0
Ma [21]	2.80	1.21	814.73	249.95	5.5
Eldeesokey [7]	2.60	1.03	829.98	233.26	4.75
Jaritz [16]	2.17	0.95	917.64	234.81	3.0
Ours	2.12	0.86	836.00	205.40	1.5

Table 1. Quantitative results on the supervised KITTI depth completion benchmark. Our method achieves state of the art performance in three metrics, iRMSE, iMAE, and MAE. [21] performs better than us by 2.6% on the RMSE metric; however, we outperform [21] on all other metrics by 24.3%, 28.9% and 17.8% on the iRMSE, iMAE and MAE, respectively. The last column is the average rank over ranks on all the four metrics.

$\mathbb{R}_+^{3 \times 3}$ centered at x in I and I' , respectively, to measure their local structural similarity. Higher scores denote more similarity; hence we can subtract the scores from 1 to form a robust version of Eq. (7). We use $P_{raw}(I'|I, d)$ and $P_{ssim}(I'|I, d)$ to represent the probability of I' given I, d measured in raw photometric value and SSIM score respectively. When the stereo pair is available, we can form the conditional prior as follows by applying conditional independence:

$$P(d|I, I', \mathcal{D}) \propto P(I'|I, d, \mathcal{D})P(d|I, \mathcal{D}) \\ = P(I'|I, d)P_{\mathcal{D}}(d|I) \quad (8)$$

Similar to the training loss Eq. (5) for the unsupervised single image depth completion setting, we can derive the loss for the stereo setting as follows:

$$L^s(w) = - \sum_{j=1}^N \log P(d_j|I_j, I'_j, z_j, \mathcal{D}) \\ = L^u(w) + \beta \sum_{j,x} \|I_j(x) - I'_j(x + s(d_j(x)))\| \quad (9)$$

where $d_j = \phi(z_j, I_j; w)$ and L^u is the loss defined in Eq. (5). Note that, the above summation term is the instantiation for $P_{raw}(I'|I, d)$, which can also be replaced by the SSIM counterpart. Rather than choosing one or the other, we compose the two with tunable parameters β_c and β_s , our final loss for stereo setting depth completion is:

$$L^s(w) = L^u(w) + \beta_c \psi_c + \beta_s \psi_s \quad (10)$$

with ψ_c represents the raw intensity summation term in Eq. (9), and ψ_s for the SSIM counterpart. Next, we elaborate our implementation details and evaluate the performance of our proposed method in different depth completion settings.

Loss	Validation Set			Test Set		
	RMSE	MAE	iRMSE	iMAE	RMSE	MAE
Ma [21]	1384.85	358.92	4.07	1.57	1299.85	350.32
L^u	1325.79	355.86	3.69	1.37	1285.14	353.16
$L^s(\psi_c)$	1320.26	353.24	3.63	1.34	1274.65	349.88
$L^s(\psi_c, \psi_s)$	1310.03	347.17	3.58	1.32	1263.19	343.46

Table 2. Quantitative results on the unsupervised KITTI depth completion benchmark. Our baseline approach using CPN as a regularizer outperforms [21] on the iRMSE, iMAE and RMSE metrics on the test set, whereas [21] marginally performs better than us on MAE by 0.8%. We note that [21] achieves this performance using photometric supervision. When including our photometric term (Eq. (10)), we outperform [21] on every metric and achieve state-of-the-art performance.

4. Implementation Details

4.1. Network architecture

We modify the public implementation of CPN [36] by replacing the input of the encoding branch with a dense depth map. Fusion of the two branches is simply a concatenation of the encodings. The encoders have only convolutional layers, while the decoder is made of transposed convolutional layers for upsampling.

Our proposed network, unlike the base CPN, as seen in Fig. 2-A, contains skip connections between the layers of the depth encoder and the corresponding decoder layers, which makes the network symmetric. We also use ResNet blocks [14] in the encoders instead of pure convolutions. A stride of 2 is used for downsampling in the encoder and the number of channels in the feature map after each encoding layer is $[64 * k, 128 * k, 256 * k, 512 * k, 512 * k]$. In all our experiments, we use $k = 0.25$ for the depth branch, and $k = 0.75$ for the image branch, taking into consideration that an RGB image has three channels while depth map only has one channel. Our network has fewer parameters than those based on early fusion (e.g. [21] used $\approx 27.8\text{M}$ parameters in total; where as we only use $\approx 18.8\text{M}$). We provide an example comparing our network architecture and that of [21] in the Supplementary Materials [37].

4.2. Training Procedure

We begin by detailing the training procedure for CPN. Once learned, we apply CPN as part of our training loss and do not need it during inference. In order to learn the conditional prior of the dense depth maps given an image, we require a dataset with images and corresponding dense depth maps. We are unaware of any real-world dataset for outdoor scenes that meets our criterion. Therefore, we train the CPN using the Virtual KITTI dataset [11]. It contains 50 high-resolution monocular videos with a total of 21, 260 frames, together with ground truth dense depth maps, generated from five different virtual worlds under different lighting and weather conditions. The original Virtual KITTI im-

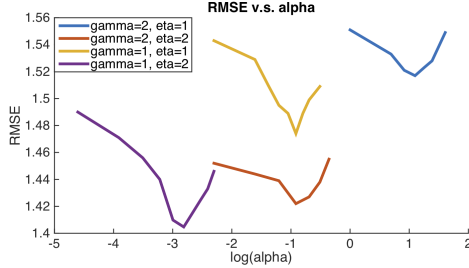


Figure 3. This plot shows the empirical study on the choice of norms γ, η in the likelihood term and the conditional prior term respectively. Each curve is generated by varying α in Eq. (5) with fixed γ, η . And the performance is measured in RMSE.

age has a large resolution of 1242×375 , which is too large to feed into a normal commercial GPU. So we crop it to 768×320 and use a batch size of 4 for training. The initial learning rate is set to $1e^{-4}$, and is halved every 50,000 steps 300,000 steps in total.

We implement our approach using TensorFlow [1]. We use Adam [17] to optimize our network with the same batch size and learning rate schedule as the training of CPN. We apply histogram equalization and also randomly crop the image to 768×320 . We additionally apply random flipping both vertical and horizontal to prevent overfitting. In the case of unsupervised training, we also perform a random shift within a 3×3 neighborhood to the sparse depth input and the corresponding validity map. We use $\alpha = 0.045$, $\beta = 1.20$ for Eq. (9), and the same α is applied with $\beta_c = 0.15$, $\beta_s = 0.425$ for Eq. (10). We choose $\gamma = 1$ and $\eta = 2$, but as one may notice in Eq. (2), the actual conditional prior also depends on the choice of the norm η . To show the reasoning behind our choice, we will present as an empirical study in Fig. 3 to show the effects of the different pairing of norms with a varying α by evaluating each model on the RMSE metric.

In the next section, we report representative experiments in both the supervised and unsupervised benchmarks.

5. Experiments

We evaluate our approach on the KITTI depth completion benchmark [29]. The dataset provides $\sim 80k$ raw image frames and corresponding sparse depth maps. The sparse depth maps are the raw output from the Velodyne lidar sensor, each with a density of about 5%. The ground truth depth map is created by accumulating the neighboring 11 raw lidar scans, with roughly 30% pixels annotated. We use the officially selected 1,000 samples for validation and we apply our method to 1,000 testing samples, with which we submit to the official KITTI website for evaluation. We additionally perform an ablation study on the effects of the sparsity of the input depth measurements on the NYUv2 indoor dataset [28] in the Supplementary Materials [37].

5.1. Norm Selection

As seen in Eq. (5), γ, η control the actual norms (penalty functions) applied to the likelihood term and conditional prior term respectively, which in turn determine how we model the distributions. General options are from the binary set $\{1, 2\}$. i.e. $\{\mathcal{L}_1, \mathcal{L}_2\}$, however, there is currently no agreement on which one is better suited for the depth completion task. [21] shows $\gamma = 2$ gives significant improvement for their network, while both [29, 16] claim to have better performance when $\gamma = 1$ is applied. In our approximation of the posterior in Eq. (5), the choice of the norms gets more complex as the modeling (norm) of the conditional prior will also depend on the likelihood model. Currently, there is no clear guidance on how to make the best choice, as it may also depend on the network structure. Here we try to explore the characteristic of different norms, at least for our network structure, by conducting an empirical study on a simple version (channel number of features reduced) of our depth completion network using different combinations of γ and η . As shown in Fig. 3, the performance on the KITTI depth completion validation set varies in a wide range with different γ, η . Clearly for our depth completion network, \mathcal{L}_1 is always better than \mathcal{L}_2 on the likelihood term. And the lowest RMSE is achieved when a \mathcal{L}_2 is also applied on the conditional prior term. Thus the best coupling is $\gamma = 1, \eta = 2$ for Eq. (5).

5.2. Supervised Depth Completion

We evaluate the proposed Depth Completion Network described in Sect. 3.1 on the KITTI depth completion benchmark. We show a quantitative comparison between our approach and the top performers on the benchmark in Tab. 1. Our approach achieves the state-of-the-art in three metrics by outperforming [7, 16], who each held the state-of-the-art in different metrics on the benchmark. We improve over [16] in iRMSE and iMAE by 2.3% and 9.5%, respectively, and [7] in MAE by 11.9%. [21] performs better on the RMSE metric by 2.6%; however, we outperform [21] by 24.3%, 28.9% and 17.8% on the iRMSE, iMAE and MAE metrics, respectively. Note in the online table of KITTI depth completion benchmark¹, all methods are solely ranked by the RMSE metric, which may not fully reflect the performance of each method. Thus we propose to rank all methods by averaging over the rank numbers on each metric, and the overall ranking is shown in the last column of Tab. 1. Not surprisingly, our depth completion network gets the smallest rank number due to its generally good performance on all metrics.

Fig. 4 shows a qualitative comparison of our method to the top performing method on the test set of the KITTI benchmark. We see that our method produces depths that

¹http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_completion

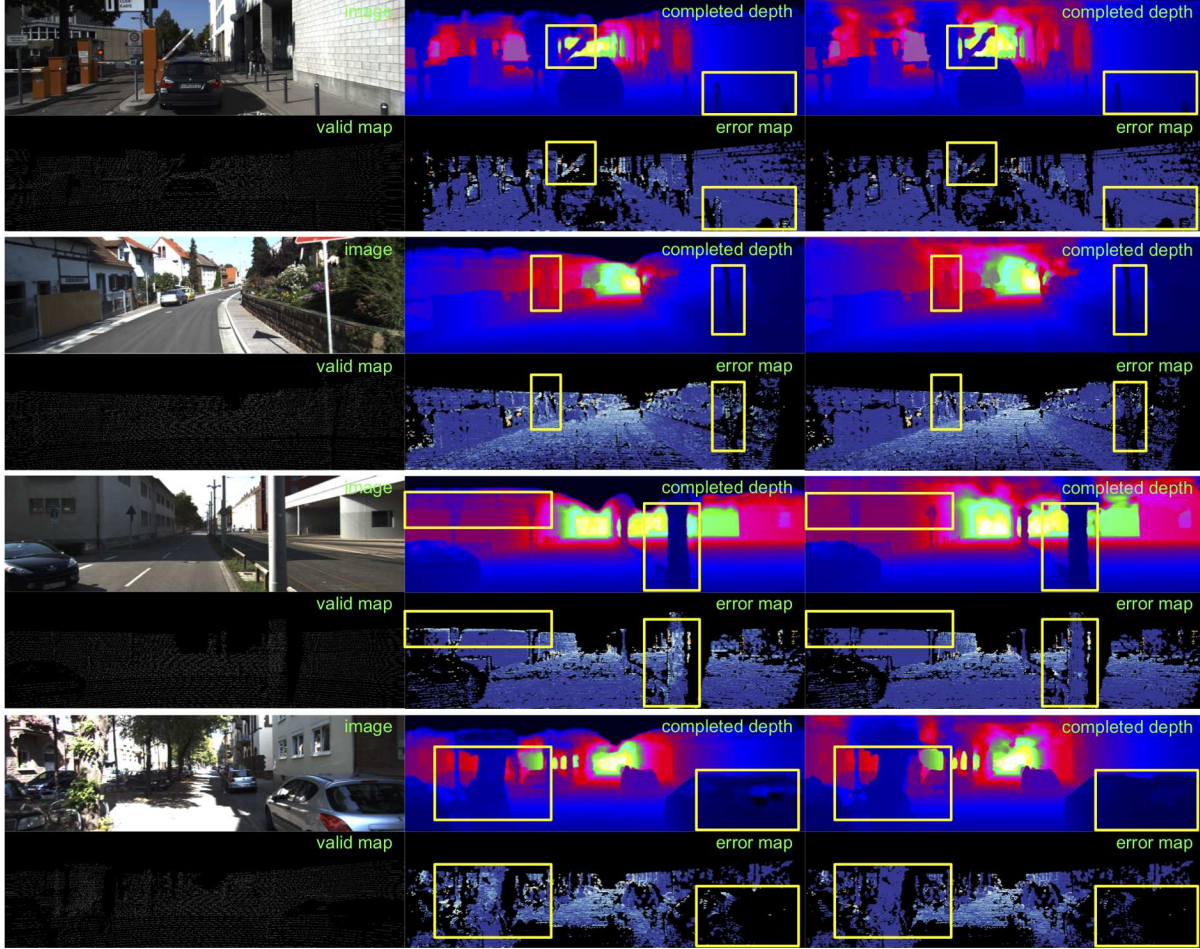


Figure 4. Qualitative comparison to Ma et al. [21] on KITTI depth completion test set in the supervised setting. Image and validity map of the sparse measurements (1st column), dense depth results and corresponding error map of [21] (2nd column) and our results and error map (3rd column). Warmer color in the error map denotes higher error. The yellow rectangles highlight the regions for detailed comparison. Note that our network consistently performs better on fine and far structures and our completed dense depth maps have less visual artifacts.

are more consistent with the scene with fewer artifacts (e.g. grid-like structures [21], holes in objects [7]). Also, our network performs consistently better on fine and far structures, which may be traffic signs and poles on the roadside that provide critical information for safe driving as shown in the second row in Fig. 4. More in the Supplementary [37].

5.3. Unsupervised Depth Completion

We show that our network can also be applied to unsupervised setting using only the training loss Eq. (5) to achieve the state-of-the-art results as well. We note that the simplest way for the network to minimize the data term is to directly copy the sparse input to the output, which will make the learning inefficient. To facilitate the training, we change the stride of the first layer from 1 to 2 and replace the final layer of the decoder with a nearest neighbor upsampling.

We show a quantitative comparison (Tab. 2) between our method and that of [21] along with an ablation study on our

loss function. We note that the results of [21] are achieved using their full model, which includes their multi-view photometric term. Our approach using just Eq. (5) is able to outperform [21] in every metric with the exception of MAE where [21] marginally beats us by 0.8%. By applying our reconstruction loss Eq. (9), we outperform [21] in every metric. Moreover, our full model Eq. (10) further improves over all other variants and is state-of-the-art in unsupervised depth completion. We present a qualitative comparison between our approach and that of [21] in Fig. 5. Visually, we observe the results of [21] still contain the artifacts as seen before. The artifacts, i.e. circles, as detailed in Fig. 5, are signs that their network is probably overfitted to the input sparse depth, due to the lack of semantic regularity. Our approach, however, does not suffer from these artifacts; instead, our predictions are globally correct and consistent with the scene geometry.

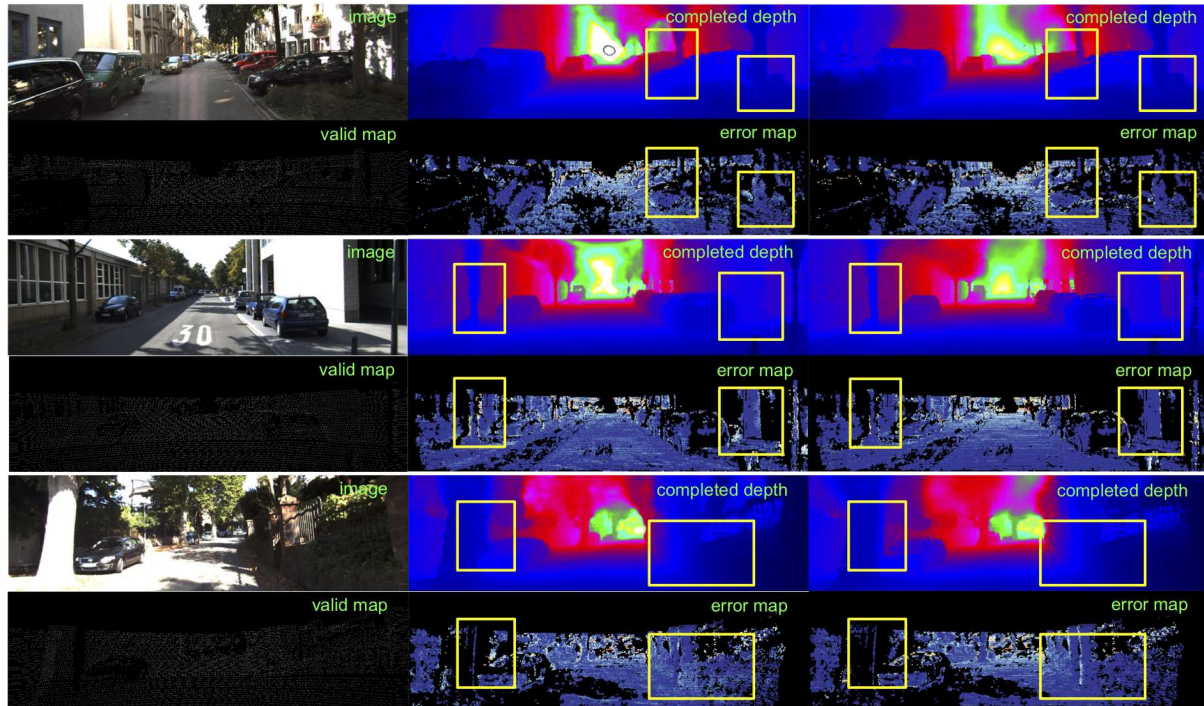


Figure 5. Qualitative comparison to Ma et al. [21] on the KITTI depth completion test set in the unsupervised setting. Image and validity map of the sparse measurements (1st column), dense depth results and corresponding error map of [21] (2nd column) and ours (3rd column). Warmer color in the error map denotes higher error. Yellow rectangles highlight the regions for detailed comparison. Note again that our network consistently performs better on fine and far structures and our completed dense depth maps have less visual artifacts (this includes the circle in the center of their prediction, row 1, column 2).

6. Discussion

In this work, we have described a system to infer a posterior probability over the depth of points in the scene corresponding to each pixel, given an image and a sparse aligned point cloud. Our method leverages a Conditional Prior Network, that allows the association of a probability to each depth value based on a single image, and combines it with a likelihood term for sparse depth measurements. Moreover, we exploit the availability of stereo imagery in constructing a photometric reconstruction term that further constrains the predicted depth to adhere to the scene geometry.

We have tested the approach both in a supervised and unsupervised setting. It should be noted that the difference between “supervised” and “unsupervised” in the KITTI benchmark is more quantitative than qualitative: the former has about 30% coverage in depth measurements, the latter about 5%. We show in Tab. 1 and 2 that our method achieves state-of-the-art performance in both supervised and unsupervised depth completion on the KITTI benchmark. Although we outperform other methods on score metrics that measures the deviation from the ground truth, we want to emphasize that our method does not simply produce a point estimate of depth, but provides a confidence measure, that can be used for more downstream processing,

for instance for planning, control and decision making.

We have explored the effect of various hyperparameters, and are in the process of expanding the testing to real-world environments, where there could be additional errors and uncertainty due to possible time-varying misalignment between the range sensor and the camera, or between the two cameras when stereo is available, faulty intrinsic camera calibration, and other nuisance variability inevitably present on the field that is carefully weeded out in evaluation benchmarks such as KITTI. This experimentation is a matter of years, and well beyond the scope of this paper. Here we have shown that a suitably modified Conditional Prior Network can successfully transfer knowledge from prior data, including synthetic ones, to provide context to input range values for inferring missing data. This is important for downstream processing as the context can, for instance, help differentiate whether gaps in the point cloud are free space or photometrically homogeneous obstacles, as discussed in our motivating example in Fig. 1.

Acknowledgments

Research supported by ARO W911NF-17-1-0304 and ONR N00014-19-1-2229.

References

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 6
- [2] M. Camplani and L. Salgado. Efficient spatio-temporal hole filling strategy for kinect depth maps. In *Three-dimensional image processing (3DIP) and applications II*, volume 8290, page 82900E. International Society for Optics and Photonics, 2012. 2
- [3] X. Cheng, P. Wang, and R. Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *European Conference on Computer Vision*, pages 108–125. Springer, Cham, 2018. 3, 5
- [4] N. Chodosh, C. Wang, and S. Lucey. Deep convolutional compressed sensing for lidar depth completion. *arXiv preprint arXiv:1803.08949*, 2018. 3
- [5] M. Dimitrievski, P. Veelaert, and W. Philips. Learning morphological operators for depth completion. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 450–461. Springer, 2018. 3
- [6] M. Dimitrievski, P. Veelaert, and W. Philips. Learning morphological operators for depth completion. In *Advanced Concepts for Intelligent Vision Systems*, 2018. 5
- [7] A. Eldesokey, M. Felsberg, and F. S. Khan. Propagating confidences through cnns for sparse data regression. *arXiv preprint arXiv:1805.11913*, 2018. 2, 3, 5, 6, 7
- [8] X. Fei, A. Wong, and S. Soatto. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. 3
- [9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3
- [10] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016. 3
- [11] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 5
- [12] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016. 3
- [13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 3
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5
- [15] Z. Huang, J. Fan, S. Yi, X. Wang, and H. Li. Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *arXiv preprint arXiv:1808.08685*, 2018. 2, 3, 4, 5
- [16] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi. Sparse and dense data with cnns: Depth completion and semantic segmentation. In *2018 International Conference on 3D Vision (3DV)*, pages 52–60. IEEE, 2018. 3, 4, 5, 6
- [17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [18] J. Ku, A. Harakeh, and S. L. Waslander. In defense of classical image processing: Fast depth completion on the cpu. *CoRR*, abs/1802.00036, 2018. 2
- [19] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 3
- [20] S. Lu, X. Ren, and F. Liu. Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3390–3397, 2014. 2
- [21] F. Ma, G. V. Cavalheiro, and S. Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018. 3, 5, 6, 7, 8
- [22] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. *arXiv preprint arXiv:1802.05522*, 2018. 3
- [23] J. Masci, J. Angulo, and J. Schmidhuber. A learning framework for morphological operators using counter-harmonic mean. In *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, pages 329–340. Springer, 2013. 3
- [24] C. Premebida, L. Garrote, A. Asvadi, A. P. Ribeiro, and U. Nunes. High-resolution lidar-based depth mapping using bilateral filter. *arXiv preprint arXiv:1606.05614*, 2016. 2
- [25] M. Ren, A. Pokrovsky, B. Yang, and R. Urtasun. Sbnets: Sparse blocks network for fast inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8711–8720, 2018. 2, 3
- [26] N. Schneider, L. Schneider, P. Pinggera, U. Franke, M. Pollefeys, and C. Stiller. Semantically guided depth upsampling. In *German Conference on Pattern Recognition*, pages 37–48. Springer, 2016. 3
- [27] J. Shen and S.-C. S. Cheung. Layer depth denoising and completion for structured-light rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1187–1194, 2013. 2
- [28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 6
- [29] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant cnns. *2017 International Conference on 3D Vision (3DV)*, pages 11–20, 2017. 2, 3, 4, 6
- [30] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 3

- [31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4
- [32] A. Wong, B.-W. Hong, and S. Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. *arXiv preprint arXiv:1903.07309*, 2019. 3
- [33] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016. 3
- [34] Y. Yang, Z. Lu, and G. Sundaramoorthi. Coarse-to-fine region selection and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2015. 3
- [35] Y. Yang and S. Soatto. S2f: Slow-to-fast interpolator flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2087–2096, 2017. 3
- [36] Y. Yang and S. Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018. 2, 3, 5
- [37] Y. Yang, A. Wong, and S. Soatto. Dense depth posterior (ddp) from single image and sparse range. *arXiv preprint arXiv:1901.10034*, 2019. 2, 4, 5, 6, 7
- [38] Y. Zhang and T. Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–185, 2018. 2, 3
- [39] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 3
- [40] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. 4