

Unsupervised Depth Completion with Calibrated Backprojection Layers

Alex Wong
 UCLA Vision Lab
 alexw@cs.ucla.edu

Stefano Soatto
 UCLA Vision Lab
 soatto@cs.ucla.edu

Abstract

We propose a deep neural network architecture to infer dense depth from an image and a sparse point cloud. It is trained using a video stream and corresponding synchronized sparse point cloud, as obtained from a LIDAR or other range sensor, along with the intrinsic calibration parameters of the camera. At inference time, the calibration of the camera, which can be different than the one used for training, is fed as an input to the network along with the sparse point cloud and a single image. A Calibrated Backprojection Layer backprojects each pixel in the image to three-dimensional space using the calibration matrix and a depth feature descriptor. The resulting 3D positional encoding is concatenated with the image descriptor and the previous layer output to yield the input to the next layer of the encoder. A decoder, exploiting skip-connections, produces a dense depth map. The resulting Calibrated Backprojection Network, or KBNet, is trained without supervision by minimizing the photometric reprojection error. KBNet imputes missing depth value based on the training set, rather than on generic regularization. We test KBNet on public depth completion benchmarks, where it outperforms the state of the art by 30.5% indoor and 8.8% outdoor when the same camera is used for training and testing. When the test camera is different, the improvement reaches 62%. Code available at: <https://github.com/alexwlwong/calibrated-backprojection-network>.

1. Introduction

Sensor platforms designed to enable interaction with physical space often include optical as well as range sensors. From cars to phones, cameras are paired with active sensors such as LIDARs, Sonars or Radars. We address the case of a single camera and a single sensor that returns the three-dimensional (3D) coordinates of a number of points far fewer than the number of pixels in the RGB image. The range sensor alone provides a sparse estimate of the Euclidean geometry of the surrounding environment, but often insufficient for planning in applications such as autonomous

navigation or manipulation. We wish to leverage the complementarity of the optical and range modalities to provide a *dense* depth map, whereby a range value¹ is associated to every pixel in the image (in the millions) as opposed to just the LIDAR or radar returns (in the thousands).

Depth completion consists of mapping a single RGB image and a sparse 3D point cloud onto a dense depth map, which requires inferring a depth value where missing. This can be done by means of regularization, or inductively using previously observed data for scenes other than the present. We assume we have available a *training set* consisting of monocular videos, corresponding sparse 3D point cloud, and intrinsic calibration matrix of the camera used for capture,² but without any manual annotation or ground-truth dense depth i.e. *unsupervised*.

Our goal is to use the training set to learn a function that, for a scene and camera not used for training, can map a sparse point cloud registered to an image, along with the matrix of intrinsic calibration parameters of the camera, and produce a dense depth map associated with the test image.

We propose a novel deep neural network architecture that leverages a *sparse-to-dense* (S2D) module and *calibrated backprojection* (KB) layers. S2D is comprised of various pooling and convolutional layers to yield a dense representation of the sparse points. A KB layer then maps camera intrinsics, input image, and current depth estimate onto the 3D scene. This can be thought of as a form of *spatial (Euclidean) positional encoding* of the image. Unlike previous architectures, camera intrinsics are an *input* to our model, as opposed to a fixed set of parameters in the training loss. This allows us more flexibility to transfer the trained model to sensor platforms other than that used for training.

Our network is trained unsupervised with the standard Photometric Euclidean Reprojection Loss (PERL) i.e. the absolute difference between a reconstructed image and the

¹The depth associated with the pixel is the Euclidean distance of the closest point in the scene along the projection ray through that pixel and the optical center. We assume the sensors to be calibrated and synchronized, and in particular the intrinsic calibration matrix of the camera is known so that pixel coordinates can be converted to Euclidean 3D coordinates.

²Typically, range and optical sensors are calibrated mechanically and pre-registered, so extrinsic calibration is not needed.

actual image measured at a time instant. We also penalize the reconstruction error of the input sparse points and Total Variation of the estimated depth map, a standard sparsity-inducing prior to reduce the penalty for large depth changes at adjacent pixels that straddle occluding boundaries. At test time, no video is necessary and inference is performed on each image and sparse point cloud independently.

These innovations allow us to improve the baseline [47] and state of the art [45] by an average of 13.7% and 8.8%, respectively, on outdoors (KITTI [41]), and 51.7% and 30.5% indoors (VOID [47]), when calibration is the same for training and testing. When different calibrations are used, our method generalizes better than the baseline and state of the art by 83% and 62%, respectively, in relative error. All of this is achieved with a smaller computational footprint thanks to the inductive bias induced by KB layers, which allows us to use a smaller network than current methods.

1.1. Related Work and Contributions

Depth completion is a form of imputation, which requires regularization that hinges the assumption that “*nearby points*” should be assigned “similar” (depth) values. Methods differ in the choice of topology i.e. what points should be considered “nearby,” and how to combine the values of such points to impute the missing depth value.

Generic Image-Based Regularization. In image topology, nearby points correspond to adjacent pixels. This is not a good choice, for their depths can be arbitrarily different at occluding boundaries. In image segmentation, the RGB values are used to define a topology to partition the image domain into connected regions of nearby points, putatively corresponding to “objects.” The topology induced by (color) values can be exploited by minimizing Total Variation (TV [36] and “Color TV” [1]) while trying to reproduce the image itself. We adopt TV as a generic regularizer since the statistics of natural range images are very similar to that of natural (intensity) images [29], whereby the gradient distribution is highly kurtotic, corresponding to homogeneous smooth regions separated by sharp boundaries.

Data-driven Regularization. “Closeness” among pixels can be defined not just within the same image, but across different images in the training set. In this case, the regularity criterion is not explicit, but implicit in the inductive bias used for training. Before training starts, the bias is encoded in the training loss (L^1 prediction error), the generic regularizers (TV), the training set, and the choice of architecture and optimization. After training is completed, all these biases are burnished in the parameters (weights) of the trained model, which inform the prediction of our depth map and therefore act as a regularizing mechanism.

Among data-driven methods for depth completion, many are **supervised**. Early works cast depth completion as com-

pressive sensing [6] and as morphological operators [7]. Recent works focused on network operations [9, 19] and architectures [3, 26, 41, 50] to effectively deal with the sparse inputs. [26] proposed an early fusion architecture while [20, 50] used late fusion to process each data modality separately. [19] performed joint concatenation and convolution to upsample the sparse depth. [3] proposed a 2D-3D fusion network while [24] used a cascade hourglass network. [4] used a convolutional spatial propagation network and [30] leveraged non-local spatial propagation. Whereas, [9, 8, 33, 34] learned uncertainty of estimates, [42] leveraged confidence maps, and [32, 49, 51] used surface normals for guidance. Like us, [28, 37, 53] proposed lightweight networks suitable for use with SLAM/VIO systems.

All of these methods require ground truth for training, which is often unavailable and, when available, prohibitively expensive [41]. Hence, these methods are limited to offline training. But even if ground truth were available online, most of these methods employ complex architectures with many layers and parameters, e.g. 25.84M for [30], 53.4M [32], and 28.99M [49], and thus are not suitable for learning online. Instead, we propose to learn dense depth from the virtually limitless amount of un-annotated images and sparse point clouds via a predictive cross-modal validation criterion. Our proposed architecture only uses 6.9M parameters and our choice of supervision allows us to continuously learn even after the system is deployed.

Unsupervised/Self-supervised depth completion assumes stereo images or monocular videos to be available during training. Both stereo [39, 50] and monocular [26, 45, 46, 47] training paradigms leverage sparse depth reconstruction and photometric reprojection error as a training signal by minimizing photometric discrepancies between the input image and its reconstruction from other views. [26] used Perspective-n-Point [23] and RANSAC [12] to align consecutive video frames. However, [26] does not generalize well to indoor scenes with many textureless surfaces. [50] learned a depth prior conditioned on the image by pretraining a separate network on ground truth from an additional dataset. As mentioned earlier, this is not scalable; also, using a network trained on a specific domain (e.g. outdoors) as supervision will not generalize (e.g. indoors). Unlike [50], our method does not require ground truth and is not limited to a specific domain. [25, 45] leverage additional synthetic datasets, which require dealing with sim-to-real; our method is able to achieve the state-of-the-art *without* needing access to additional data.

The challenge of depth completion is precisely the sparsity, which renders convolutions ineffective as the activations of early layers tend to be zeros as well. To obtain a denser representation, early layers must propagate (or densify) the signal. As a result, [26, 39, 50] employed very deep networks with many layers and parameters in order to

learn the map from sparse depth and image to dense depth. To handle this problem, [47] approximated the scene with a hand-crafted mesh, but it is not differentiable and prone to errors in regions with very few points or complex structures. [45] proposed spatial pyramid pooling (SPP), but their max pooling layers decimated details on closer objects. Instead, we propose a fully differentiable sparse-to-dense module that learns the trade-off between density and detail to retain both near and far structures.

Our work goes counter to the trend of forgoing inductive bias, i.e. learning everything with generic architectures like Transformers [43], including what we already know such as basic Euclidean geometry. Our model has a strong inductive bias in our calibrated backprojection layer, which incorporates the calibration matrix directly into the architecture to yield an RGB representation lifted into scene topology via 3D positional encoding. This may seem futile as we could just add intrinsics to the long list of parameters to be learned [15]. However, unlike semantic retrieval, spatial inference requires *identifiability*: There is *one* true scene in front of us, and unless information about calibration is available and properly exploited, inference yields one of infinitely many depth maps that are equally good at predicting the next frame in the training set. Since there is no supervision, calibration mediates the relation between the prediction error and the *true* depth. Because existing methods use calibration in the computation of the loss, which the intrinsics are encoded in the weights, hampering transferability. In our architecture, calibration is an input, which can be changed at inference time. While one could pre-process the images to a canonical calibration, this introduces latency, cost and artifacts that can affect the reconstruction quality. We note that [16, 35] proposed backprojection as a layer and [10] used calibration as input, but we are the first to consider an RGB 3D representation for depth completion.

Our contributions include (a) a sparse-to-dense module that learns a dense representation of the sparse point cloud, (b) an unsupervised depth completion method that takes calibration information as input to the model, and (c) incorporates it directly in the architecture through a novel *calibrated backprojection* module, which represents spatial positional encoding that is transferred laterally across different branches of the encoder. The resulting inductive bias helps select, among all depth, maps compatible with the prediction loss, those that result in a Euclidean (calibrated) reconstruction. The strong inductive bias allows us to (d) reduce the computational footprint, increase generalization and achieve performance beyond the state of the art despite having fewer parameters.

2. Method Formulation

Our goal is to recover a 3D scene from an RGB image $I : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}_+^3$ and the associated sparse point cloud

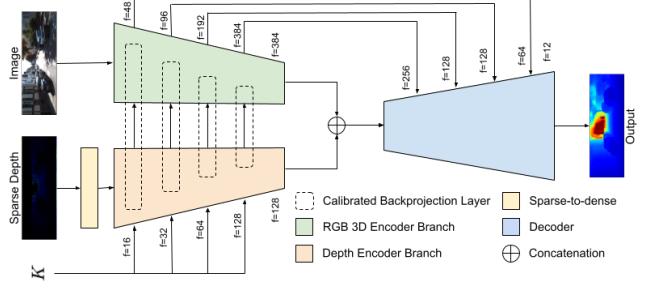


Figure 1: *KBNet architecture*. Our architecture takes, as input, an RGB image, the corresponding sparse depth map and camera calibration matrix. We first learn a dense representation of the sparse points with our sparse-to-dense module. The result of which and the calibration matrix are used for calibrated lifting, which allows us to backproject image features onto 3D space (akin to spatial positional encodings) to yield a RGB 3D representation. Our network is very light-weight and fast, yet achieves the state of the art.

projected onto the image plane $z : \Omega_z \subset \Omega \mapsto \mathbb{R}_+$, without access to ground-truth depth annotations.

We propose a sparse-to-dense module (Fig. 2) f_ω , parameterized by ω , that captures local and global structure of the sparse inputs by combining min and max pooling at different scales. The result is a dense or quasi-dense depth representation $f_\omega(z)$, depending on the sparsity of the input, which frees the rest of network to utilize its early convolutional layers to learn scene structure rather than to densify the input – making the overall architecture more efficient.

The sparse-to-dense module (Sec. 2.1) is part of an overall encoder-decoder architecture f_θ , parameterized by θ , called KBNet (Sec. 2.2), that includes a Calibrated Backprojection layer which explicitly backprojects pixels onto 3D space using intrinsic camera calibration and depth encoding from f_ω . Unlike previous works [26, 39, 45, 47, 50] that encode depth and image in two separate branches, we leverage camera calibration and our depth encoding to lift the image representation to 3D and passed it to the decoder via skip connections. KBNet (Fig. 1) produces dense depth $\hat{d} := f_\theta(f_\omega(z), I, K)$, where $K \in \mathbb{R}^{3 \times 3}$ is the upper-triangular matrix of intrinsic calibration parameters. To train our model, we use monocular videos to compose a loss function from temporally adjacent frames (Sec. 2.3).

2.1. Sparse-to-Dense Module (S2D)

Our S2D module f_ω (Fig. 2) performs multi-scale densification on the input sparse depth map z using a series of min and max pooling layers with various kernel sizes, which are chosen based on the sparsity of the point cloud e.g. from LIDAR returns or sparse points tracked by VIO [11] (see Supp. Mat. for kernel sizes). The outputs of the pooling layers are concatenated and fed into three 1×1 convolutions to learn the trade-offs between pooling types and

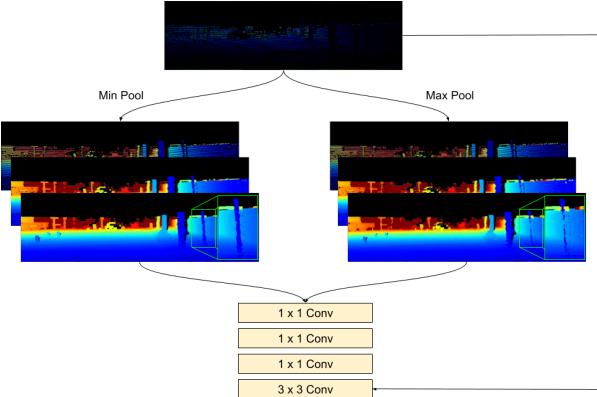


Figure 2: *Sparse-to-dense module*. We perform min and max pooling with various kernel sizes to produce a dense representation. There exists trade-offs between density and detail (large vs. small kernel sizes) and preservation of near and far structures (min vs. max pooling, as highlighted in green). We balance these trade-offs with 1×1 convolutions and fuse the result with the input via a 3×3 convolution.

kernel sizes. The result of which is fused with the z via a 3×3 convolutional layer, yielding a dense or quasi-dense depth representation that is fed to the rest of the network f_θ .

Because the depth inputs are sparse, we design our min pooling layers to avoid pooling zeros or invalid (negative) depth values. We set all values $z(x)$ less than zero to be infinity for $x \in \Omega$:

$$z'(x) = \begin{cases} z(x) & \text{if } z(x) > 0 \\ \infty & \text{otherwise.} \end{cases} \quad (1)$$

z' is fed to a min pooling layer with $k \times k$ kernel size,

$$p = \text{minpool}(z', k). \quad (2)$$

Finally, for all x , any infinity values pooled due to large empty regions are set to zero:

$$p_{\text{min}}(x) = \begin{cases} p(x) & \text{if } p(x) \neq \infty \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Our approach involves two main trade-offs: (i) density versus detail and (ii) preservation of near versus far structures.

Density versus details. For the purpose of densification, one may perform pooling with large kernel sizes, but it comes at the expense of details of local structures. In contrast, pooling with small kernel sizes in an attempt to retain detail will result in very few neuron activations, which hinders learning. Hence, to retain local details while obtaining a dense representation, we propose to perform pooling with both small and large kernel sizes.

Near versus far. When pooled solely with max pooling, farther structures are preserved, but details of the closer

ones are decimated as the kernel size grows larger. For instance in Fig. 2, thin structures close to the camera i.e. the highlighted pole “disappears” due to large max pooling kernel size. On the other hand, when only using min pooling, the closer structures become more prominent, but in turn, the farther regions are corrupted. Moreover, in cluttered scenes, min pooling causes adjacent structures to “bleed” into each other. Hence, to preserve close and far structures, we employ both min and max pooling layers.

To optimize for both trade-offs, we concatenate the outputs of min and max pooling together and feed them into 1×1 convolutional layers. Finally, we use a 3×3 convolution to fuse the multi-scale depth features back into the original sparse depth via a residual connection, yielding a dense representation $f_\omega(z)$ to be fed to f_θ .

We note that our S2D bears some resemblance to spatial pyramid pooling (SPP) [18]; however, SPP was designed to ensure the same size feature maps are maintained when different size of inputs. It is also intended to operate on dense inputs. While [45] also proposed an SPP for sparse inputs, its use of max pooling decimated details for nearby structures. Neither are substitutes for our S2D module.

2.2. KBNet Architecture

Motivation. Unsupervised methods [26, 45, 46, 47] use the photometric reprojection error ℓ_{perl} as a training signal. The input image I_t is reconstructed from temporally adjacent frames I_τ for $\tau \in T \doteq \{t-1, t+1\}$ to yield \hat{I}_τ ,

$$\hat{I}_\tau(x, \hat{d}, g_{\tau t}) = I_\tau(\pi g_{\tau t} K^{-1} \bar{x} \hat{d}(x)), \quad (4)$$

and the per pixel photometric reprojection error is measured by $\ell_{\text{perl}} = |\hat{I}_\tau(x, \hat{d}, g_{\tau t}) - I_t(x)|$. Here $\bar{x} = [x^\top, 1]^\top$ are the homogeneous coordinates of $x \in \Omega$. Using the notation in [27], $g_{\tau t} \in SE(3)$ is the relative pose (rotation and translation) of the camera from time t to time τ , K denotes the intrinsic calibration matrix, and π is a canonical perspective projection. For simplicity, we will refer to the reconstruction from time τ at a coordinate x as $\hat{I}_\tau(x)$.

Inferring Euclidean structure and motion in the absence of calibration information is notoriously difficult and dependent on conditions rarely satisfied in ordinary training videos, such as rotation around three independent axes [27]. Minimizing any form of ℓ_{perl} forces the network to implicitly learn the calibration matrix K , as all prior work does. As pretrained models are commonly deployed on sensor platforms different than those used during training, this hinders generalization as the network becomes overfitted to the camera used to collect training data. In contrast, our network, KBNet, takes it as input; this allows us to use different calibrations in training and test, which significantly improves generalization (Table 5).

Calibrated Backprojection Layers take, as input, the depth and RGB image encodings, and the camera calibra-

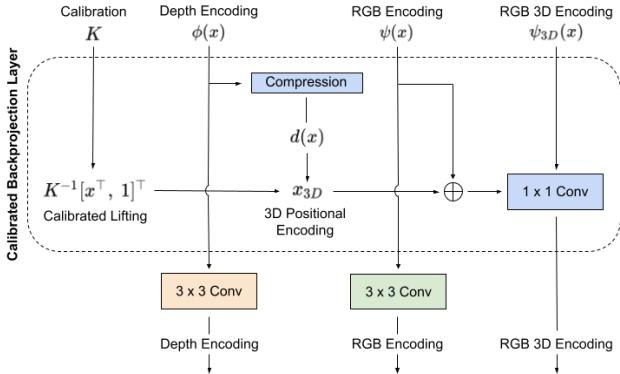


Figure 3: *Calibrated Backprojection (KB) Layer*. The standard depth and color image encoding layers [47] are combined using the calibration matrix as additional input. Calibration is used to lift pixel coordinates to three dimensions, which are backprojected by a compressed depth descriptor into a 3D positional encoding. The result is concatenated with the image encoding and the output of the previous KB layer, and fused with a 1×1 convolution. This yields an RGB 3D representation, which is used as a skip connection to the decoder and input to subsequent layers.

tion matrix K and output not only the corresponding encodings of the depth map and of the RGB image, but also an encoding of the RGB image backprojected onto 3D space. Once we have formed this RGB 3D representation, it is fed as input to subsequent Calibrated Backprojection (KB) layers and as skip connection to the decoder and once we have form this representation (Fig. 3).

To realize a KB layer, first, we use the calibration matrix to lift the coordinates of each pixel $x \in \Omega$ to three dimensional space $x \rightarrow K^{-1}\bar{x}$. Then, the feature map of the depth encoder $\phi(x) \in \mathbb{R}^M$, with M ranging from 16 in the first layer to 128 in the last one, is collapsed to a scalar by a trainable projection or ‘‘compression’’ module q , $d(x) = q^\top \phi(x)$. The imputed depth $d(x)$ is used to back-project the lifted coordinate \bar{x} to yield a 3D positional encoding for each pixel $x_{3D} = K^{-1}\bar{x}d(x)$.

Here $\Omega \subset \mathbb{R}^2$ is discretized into a lattice of $H \times W$ pixels in the first layer, corresponding to the resolution of the original image, that decreases by a factor of 2 in each subsequent layer until the 5-th or last layer at $H/32 \times W/32$. Hence, the intrinsics parameters, focal lengths and principal point, must also be scaled by the same factor according to the resolution reduction in each layer.

The 3D positional encoding is concatenated with the image encoding $\psi(x) \in \mathbb{R}^N$, and, if available, the output of the previous KB layer $\psi_{3D}(x) \in \mathbb{R}^N$ where N ranges from 48 in the first layer to 386 in the last. This is fused together by a 1×1 convolution to yield the output RGB 3D encoding. This encoding is fed to the next layer and also replaces the typical RGB skip connection to the decoder. Finally, the

output depth and image encodings of the KB layer are produced by convolving separate 3×3 kernels. After which, both are also passed to the next layer as input.

In addition to benefits of generalization (Table 5), KB layers also produce depth estimates that better respect object boundaries. Because each layer encodes ‘‘closeness’’ based on the scene topology via 3D positional encoding rather than the 2D image topology (as in previous works), adjacent pixels in the image that are often confused to be close are now well separated (Fig. 4) and hence distinct adjacent objects are better delineated and points belonging to the same surface are better regularized. This reduces the common bleed effect observed when a depth map is back-projected to a point cloud in 3D. Moreover, by instilling 3D structure as an architectural inductive bias, we enable a faster and slimmer network with fewer layers and parameters to achieve better performance (see Table 2, 4).

We note that our S2D module complements our KB layers as it provides us with dense or quasi-dense depth representation. Without it, we are left with sparse geometry, which limits the potential performance gain. Yet, as demonstrated in Table 3, there are still benefits to using calibrated backprojection with a sparse representation.

2.3. Loss Function

Similar to previous works [26, 45, 47], our loss function is the linear combination of three terms:

$$\mathcal{L} = w_{ph}\ell_{ph} + w_{sz}\ell_{sz} + w_{sm}\ell_{sm} \quad (5)$$

where ℓ_{ph} denotes photometric consistency, ℓ_{sz} sparse depth consistency, and ℓ_{sm} local smoothness. Each term is weighted by their associated w (see Sec. 3.1).

Photometric Consistency. As mentioned in Sec. 2.2, unsupervised methods leverage photometric reprojection error as a supervisory signal by reconstructing I_t from I_τ for $\tau \in T \doteq \{t-1, t+1\}$ via Eqn. 4. To accomplish this, one can obtain pose from a VIO [11] or employ a pose network to estimate the relative pose between I_t and I_τ (see full system diagram in Supp. Mat.). We note that pose is only needed for training and is not used at test time.

From the reconstructions, the photometric consistency loss measures the average photometric reprojection error using a combination of L^1 penalty and SSIM [44]:

$$\begin{aligned} \ell_{ph} &= \frac{1}{|\Omega|} \sum_{\tau \in T} \sum_{x \in \Omega} w_{co} |\hat{I}_\tau(x) - I_t(x)| + \\ &\quad w_{st} (1 - \text{SSIM}(\hat{I}_\tau(x), I_t(x))), \end{aligned} \quad (6)$$

w_{co} and w_{st} are weights for each term and are discussed in Sec. 3.1. We note that if $g_{\tau t}$ is estimated via a pose network, instead of a VIO, it can be jointly learned with KBNet (Fig. 1) as a by product from minimizing Eqn. 6 and 7, and hence does not require any extra supervision.

Metric	Definition
MAE	$\frac{1}{ \Omega } \sum_{x \in \Omega} \hat{d}(x) - d_{gt}(x) $
RMSE	$(\frac{1}{ \Omega } \sum_{x \in \Omega} \hat{d}(x) - d_{gt}(x) ^2)^{1/2}$
iMAE	$\frac{1}{ \Omega } \sum_{x \in \Omega} 1/\hat{d}(x) - 1/d_{gt}(x) $
iRMSE	$(\frac{1}{ \Omega } \sum_{x \in \Omega} 1/\hat{d}(x) - 1/d_{gt}(x) ^2)^{1/2}$

Table 1: *Error metrics.* d_{gt} denotes the ground-truth depth.

Sparse Depth Consistency. Minimizing the reprojection error will reconstruct the scene structure up to an unknown scale. To ground the predictions to *metric* scale, we minimize the L^1 difference between our predictions \hat{d} and the sparse depth inputs over its domain (Ω_z):

$$\ell_{sz} = \frac{1}{|\Omega_z|} \sum_{x \in \Omega_z} |\hat{d}(x) - z(x)|. \quad (7)$$

Local Smoothness. We enforce local smoothness and connectivity over \hat{d} by minimizing the L^1 penalty on its gradients in the x - (∂_X) and y - (∂_Y) directions. We also weight each term using its respective image gradients, $\lambda_X = e^{-|\partial_X I_t(x)|}$ and $\lambda_Y = e^{-|\partial_Y I_t(x)|}$, to allow discontinuities along object boundaries:

$$\ell_{sm} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \lambda_X(x) |\partial_X \hat{d}(x)| + \lambda_Y(x) |\partial_Y \hat{d}(x)|. \quad (8)$$

3. Experiments and Results

We evaluate our method on benchmark datasets, KITTI [41] for outdoors settings, and VOID [47] for indoors, using metrics described in Table 1. We also demonstrate that our approach generalizes well to scenes captured by camera setup different than that used to collect the training set by training our model on VOID and testing it on NYUv2 [40].

3.1. Implementation Details

We implemented our method in PyTorch [31]. End-to-end inference takes 16ms per frame. We used Adam [21] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize our network. Training on KITTI [41] takes 70 hours for 60 epochs, VOID [47] 16 hours for 15 epochs, and NYUv2 [40] 13 hours for 15 epochs on an Nvidia GTX 1080Ti GPU. We use a batch size of 8 with 768×320 crops for KITTI, 640×480 for VOID and 576×416 for NYUv2. For KITTI, we choose $w_{ph} = 1$, $w_{co} = 0.15$, $w_{st} = 0.95$, $w_{sz} = 0.6$, and $w_{sm} = 0.04$; for VOID and NYUv2, we set $w_{sz} = 2$ and $w_{sm} = 2$. For detailed learning rate schedule, augmentations and S2D kernel sizes used for each dataset, please see Supp. Mat.

3.2. Datasets

KITTI [41] provides $\approx 80,000$ raw image frames and associated sparse depth maps. The sparse depth maps are the

Method	# Param	Time	MAE	RMSE	iMAE	iRMSE
SS-S2D [26]	27.8M	80ms	350.32	1299.85	1.57	4.07
IP-Basic [22]	0	11ms	302.60	1288.46	1.29	3.78
DFuseNet [39]	n/a	80ms	429.93	1206.66	1.79	3.62
DDP* [50]	18.8M	80ms	343.46	1263.19	1.32	3.58
VOICED [47]	9.7M	44ms	299.41	1169.97	1.20	3.56
AdaFrame [46]	6.4M	40ms	291.62	1125.67	1.16	3.32
SynthProj* [25]	2.6M	60ms	280.42	1095.26	1.19	3.53
ScaffNet* [45]	7.8M	32ms	280.76	1121.93	1.15	3.30
Ours	6.9M	16ms	256.76	1069.47	1.02	2.95

Table 2: *Quantitative results on the KITTI test set.* Our method outperforms all unsupervised methods across all metrics on the KITTI leaderboard. Compared to the baseline [47], we improve by an average of 13.7% across all metrics while using 29% fewer parameters. * denotes methods that use additional synthetic data for training.

raw output from the Velodyne lidar sensor, each with a density of $\approx 5\%$. Ground-truth depth is obtained by accumulating 11 neighbouring raw lidar scans. Semi-dense depth is available for the lower 30% of the image space. We use the official 1,000 samples for validation and test on 1,000 designated samples (evaluated on their online test server).

VOID [47] contains synchronized 640×480 RGB images and sparse depth maps of indoor (laboratories, classrooms) and outdoor (gardens) scenes. ≈ 1500 sparse depth points (covering $\approx 0.5\%$ of the image) are the set of features tracked by XIVO [11], a VIO system. The ground-truth depth maps are dense and are acquired by active stereo. The entire dataset contains 56 sequences with challenging motion. Of the 56 sequences, 48 sequences ($\approx 40,000$) are designated for training and 8 for testing. The testing set contains 800 frames. We follow the evaluation protocol of [47] and cap the depths between 0.2 and 5 meters.

NYUv2 [40] consists of 372K synchronized 640×480 RGB images and depth maps for 464 indoor scenes (household, offices, commercial), captured with a Microsoft Kinect. The official split consisting in 249 training and 215 test scenes. For training, we evenly sample a subset of the training split to yield 46K frames. We use the official validation set of 795 images and test set of 654 images. Because there are no sparse depth maps provided, we sampled ≈ 1500 points from the depth map via Harris corner detector [17] to mimic the sparse depth produced by SLAM/VIO.

3.3. KITTI Depth Completion Benchmark

We compare our method against recent unsupervised depth completion methods on the KITTI test set in Table 2 (results taken from online leaderboard). Compared to the baseline [47], we improve by an average of 13.7% across metrics and by as much as 17.1% in iRMSE while reducing

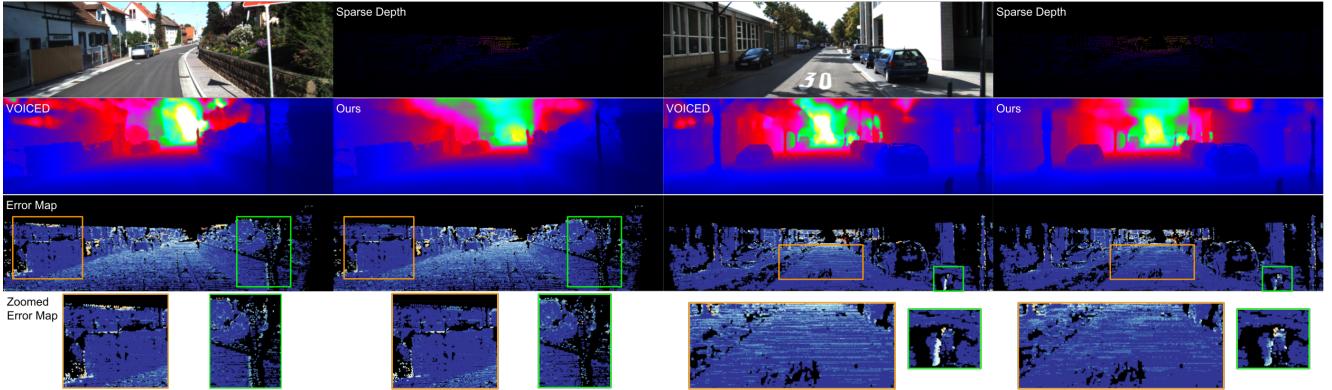


Figure 4: *Qualitative results on KITTI test set.* Head-to-head comparison against [47]. Thanks to our 3D positional encoding, our method performs well on regions where adjacent structures in 2D image space are far apart in the 3D scene e.g. street sign and wall (left panel, highlighted in green) and far region of the road (right panel, in orange).

Method	MAE	RMSE	iMAE	iRMSE
VOICED [47] w/o Scaffolding	347.14	1330.88	1.46	4.22
VOICED [47]	305.06	1239.06	1.21	3.71
Ours w/o S2D	287.76	1184.24	1.12	3.48
Ours w/o KB layers	285.97	1171.88	1.11	3.40
Ours w/ Scaffolding [47]	275.56	1183.57	1.08	3.39
Ours w/ SPP [18, 45]	273.08	1177.69	1.07	3.35
Ours	260.44	1126.85	1.03	3.20

Table 3: *Ablation study on KITTI validation set.* Without S2D (row 3), our performance degrade because our 3D positional features will only encode sparse geometry, but we still beat [47] in rows 1, 2 (“w/o Scaffolding” is [47] with sparse representation). We observe similar degradation without KB layers (row 6, replaced with VGG block used by [47]). Substituting our S2D with Scaffolding [47] or SPP [18, 45] also hurts performance (rows 7, 8).

model size by 29%. Overall, we beat the best performing method [45] by an average of 8.8% and up to 10.6% on the iMAE metric with a 11.5% reduction in model size. We note that top methods [25, 45] use *additional* synthetic data for training; whereas, we do not. Also, for inference, our method takes 16ms per image (62 FPS), which is $2.75 \times$ faster than [47]³ and $2 \times$ faster than the state of the art [45]. We note that our method significantly improves the iMAE and iRMSE metrics, to the point where we are comparable to some of the supervised methods for close range performance. For example, our iMAE score is ranked 5th across all methods (see Table 9, 10 in Supp. Mat.). To the best of our knowledge, we are the first work in unsupervised depth completion to demonstrate comparable performance to supervised methods.

³The reported run time of [47] on the KITTI leaderboard did not include their scaffolding step; whereas, the number in Table 2 accounts for it.

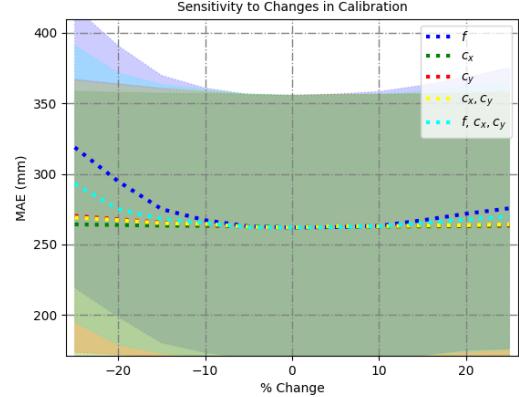


Figure 5: *Sensitivity to changes in calibration on KITTI.* Focal length and principal point are altered to test sensitivity to changes in intrinsics parameters. Our method is robust to change up to $\approx 10\%$. After which, performance degrades.

To show the improvements from our contributions, we show head-to-head qualitative comparisons against the baseline [47] in Fig. 4. Our method performs better in regions where depth discontinuities occur in image topology i.e. street sign and wall (left panel, highlighted in green) and far regions of the road (right panel, in orange). This is in part thanks to our calibrated backprojection (KB) layer which goes counter to the current trend of learning everything with generic architectures, including what we already know about basic Euclidean geometry. Our KB layers imposes strong inductive bias by incorporating the camera intrinsic calibration matrix to yield 3D positional encoding that lifts the image representation into scene topology – this delineates points where in 2D image topology are “close”, but can be far in 3D scene topology.

Table 3 shows an ablation study on the KITTI validation set. As mentioned in Sec. 2.2, our sparse-to-dense module (S2D) provides dense depth representation which in turn enables dense 3D topology in our calibrated backprojection

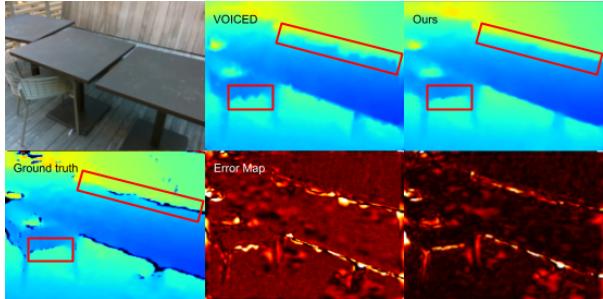


Figure 6: Qualitative results on VOID test set. Comparison against [47]. Our method performs better overall.

Method	# Param	Time	MAE	RMSE	iMAE	iRMSE
SS-S2D [26]	27.8M	59ms	178.85	243.84	80.12	107.69
DDP [50]	18.8M	54ms	151.86	222.36	74.59	112.36
VOICED [47]	9.7M	29ms	85.05	169.79	48.92	104.02
ScaffNet [45]	7.8M	25ms	59.53	119.14	35.72	68.36
Ours	6.9M	13ms	39.80	95.86	21.16	49.72

Table 4: Quantitative results on VOID test set. We outperform all competing methods across all metrics. Compared to [45], we improve by an average of 30.5%.

(KB) layers. Hence, removing it (“w/o S2D) will hurt performance because it results in a *sparse* 3D positional encoding. Nonetheless, sparse geometry is still helpful as we outperform [47] in rows 1, 2. Similarly, replacing our KB (“w/o KB layers”) with VGG blocks used by [47] also hurts performance as the model now lacks 3D spatial position. We show in rows 5 and 6 that one cannot simply substitute S2D with scaffolding [47] or SPP [18, 45].

In Fig. 10, we perform a sensitivity study of our model to calibration on the KITTI validation set. To this end, we altered the calibration by increasing or decreasing focal length (f) and/or principal point (c_x, c_y) and feed it as input. Our model is robust to changes up to $\approx 10\%$; after which, performance degrades. While changes in c_x, c_y have minor effects (which is scene-dependent), we observe a sharp decrease in performance when we decrease f by 20 to 25%. This is because, geometrically, decreasing f backprojects points to a larger field of view, distorting surfaces and sending points of the same surface far from each other. Increasing f conversely “packs” them tighter; this is okay for small increases, but for larger values, points will get “squashed together” – thus hurting performance. Also, to quantify the effect of sparsity, we provide a sensitivity study on various density levels in Supp. Mat.

3.4. VOID Depth Completion Benchmark

In the indoor scenario, the point clouds are on orders of hundreds to several thousand points (if we are being generous); hence, because of the sparsity, perturbations to the point cloud can yield vastly different sparse geometry.

Method	Trained on	MAE	RMSE	iMAE	iRMSE
VOICED [47]	NYUv2	127.61	228.38	28.89	54.70
VOICED [47]	VOID	178.87	329.28	42.57	105.93
ScaffNet [45]	NYUv2	117.49	199.31	24.89	44.06
ScaffNet [45]	VOID	155.20	241.42	31.77	52.62
Ours	NYUv2	105.76	197.77	21.37	42.74
Ours	VOID	117.18	218.67	23.01	47.96

Table 5: Quantitative results on the NYUv2 test set. Column titled “Trained on” denotes the dataset each method is trained on. [45, 47] degrade much more than our method when tested on a dataset captured by a different sensor platform than the one used for gathering its training data.

This increases a model’s sensitivity to the distribution to the sparse points. As there exists many complex scene layouts for the indoor setting, learning a dense representation and understanding the 3D topology of the scene become even more important. This is shown in Table 4 where we outperform [26, 45, 47, 50] across all metrics to achieve the state of the art on VOID. A key comparison is between our method and [47]. Even though [47] creates a hand-crafted scaffolding of the scene to obtain a dense representation, because there are very few points, it is prone to error i.e. forming surfaces between discontinuous objects and sensitive to changes in the points sampled. This is where our method shines. By optimizing for the trade-off between density and detail, our S2D module learns to exploit the natural statistics of the dataset to obtain a dense representation more compatible with the scene. Also, our KB layers introduces 3D topology as an inductive bias, allowing the network to delineate points that are close in image topology, but are far in scene topology – culminating in 51.7% and 30.5% improvement over [47] and the state of the art [45], respectively.

In Table 5, we show that our method generalizes well to sensor platforms not used in the training set by training our method on VOID (captured on Intel RealSense) and testing it on NYUv2 (Microsoft Kinect). Similarly, we test models pretrained on VOID released by [45, 47] on NYUv2. We also train our method and [45, 47] from scratch on NYUv2 to show the paragon performance (rows 1, 3, 5). Rows 1 shows that [47] does not generalize well to NYUv2 where error increases by 56% (as much as 94% in iRMSE). While [45] does better, there is still a sharp decrease of 25.1% in performance. This is in part due to the change in sensor platform as well scene distribution in NYUv2. While we do not achieve paragon performance, our method generalizes better with a reasonable 10% increase in error – improving over [47] by 83% and [45] by 62% in relative error. We note that while training on the full set for NYUv2 should yield better results for paragon performance, our model trained

on VOID performs better than VOICED [47] and comparable to ScaffNet [47] trained on the subset of NYUv2. For qualitative comparisons, please see Fig. 14 in Supp. Mat.

4. Discussion

We present an approach to unsupervised depth completion that imposes strong inductive biases on Euclidean reconstruction in the architecture, rather than learning from data with a generic model such as a Transformer. This presents some advantages. First, it allows feeding calibration as an input, which means that we can easily use a model trained with a certain sensor platform with a different one at inference time. Second, the calibrated backprojection layer explicitly incorporates a basic geometric image formation model based on Euclidean transformations in 3D and central perspective projection onto 2D. This allows us to reduce the model size while still achieving the state of the art.

However, imposing strong inductive biases also presents some risks and limitations. First, if the camera is miscalibrated, inputting the wrong calibration can backfire, yielding distorted depth maps. Second, only a very rudimentary calibration model is used, so if a sensor platform has fancy optics such as omnidirectional lenses, one cannot use one of our pre-trained models but rather has to modify the core backprojection module. Third, even with these ad-hoc architectural choices, our model suffers the limitations of all imputations, which is that where there is insufficient evidence to constrain the solution, the regularizer dominates, which is a form of hallucination and can yield wildly wrong inferences. This would be mitigated by having an accurate measure of uncertainty associated to the depth map, this is an open problem well beyond our focus here.

Acknowledgements. This work was supported by ARL W911NF-20-1-0158 and ONR N00014-17-1-2072.

References

- [1] Peter Blomgren and Tony F Chan. Color tv: total variation methods for restoration of vector-valued images. *IEEE transactions on image processing*, 7(3):304–309, 1998. [2](#)
- [2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. [14](#)
- [3] Yun Chen, Bin Yang, Ming Liang, and Raquel Urtasun. Learning joint 2d-3d representations for depth completion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10023–10032, 2019. [2, 18](#)
- [4] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10615–10622, 2020. [2, 18](#)
- [5] 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. [18, 19](#)
- [6] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. In *Asian Conference on Computer Vision*, pages 499–513. Springer, 2018. [2, 18, 19](#)
- [7] Martin Dimitrovski, Peter Veelaert, and Wilfried Philips. Learning morphological operators for depth completion. In *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2018. [2, 18, 19](#)
- [8] Abdelrahman Elde索key, Michael Felsberg, Karl Holmquist, and Michael Persson. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12014–12023, 2020. [2](#)
- [9] Abdelrahman Elde索key, Michael Felsberg, and Fahad Shahbaz Khan. Propagating confidences through cnns for sparse data regression. In *Proceedings of British Machine Vision Conference (BMVC)*, 2018. [2](#)
- [10] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. [3](#)
- [11] Xiaohan Fei, Alex Wong, and Stefano Soatto. Geosupervised visual depth prediction. *IEEE Robotics and Automation Letters*, 4(2):1661–1668, 2019. [3, 5, 6, 13, 19](#)
- [12] Martin A Fischler and Robert 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. [2](#)
- [13] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. [18, 19](#)
- [14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. [13](#)
- [15] 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/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. [3](#)
- [16] Ankur Handa, Michael Bloesch, Viorica Pătrăucean, Simon Stent, John McCormac, and Andrew Davison. gvn: Neural network library for geometric computer vision. In *European Conference on Computer Vision*. Springer, 2016. [3](#)
- [17] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. [6](#)
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. [4, 7, 8, 14](#)
- [19] Zixuan Huang, Junming Fan, Shenggan Cheng, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Hms-net: Hierarchical

- multi-scale sparsity-invariant network for sparse depth completion. *IEEE Transactions on Image Processing*, 29:3429–3441, 2019. 2
- [20] Maximilian Jaritz, Raoul De Charette, Emilie Wirbel, Xavier Perrotton, and Fawzi 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. 2
- [21] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*, 2015. 6, 13
- [22] Jason Ku, Ali Harakeh, and Steven L Waslander. In defense of classical image processing: Fast depth completion on the cpu. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 16–22. IEEE, 2018. 6, 18
- [23] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 2
- [24] Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, Chong Zhang, et al. A multi-scale guided cascade hourglass network for depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–40, 2020. 2
- [25] Adrian Lopez-Rodriguez, Benjamin Busam, and Krystian Mikolajczyk. Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 6, 7, 18, 19
- [26] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *International Conference on Robotics and Automation (ICRA)*, pages 3288–3295. IEEE, 2019. 2, 3, 4, 5, 6, 8, 14, 18, 19
- [27] Yi Ma, Stefano Soatto, Jana Kosecka, and S Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*, volume 26. Springer Science & Business Media, 2012. 4
- [28] Nathaniel Merrill, Patrick Geneva, and Guoquan Huang. Robust monocular visual-inertial depth completion for embedded systems. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021. 2
- [29] David Bryant Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989. 2
- [30] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In-So Kweon. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision, ECCV 2020*. European Conference on Computer Vision, 2020. 2, 18, 19
- [31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 6
- [32] 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. 2, 18, 19
- [33] Chao Qu, Wenxin Liu, and Camillo J Taylor. Bayesian deep basis fitting for depth completion with uncertainty. *arXiv preprint arXiv:2103.15254*, 2021. 2
- [34] Chao Qu, Ty Nguyen, and Camillo Taylor. Depth completion via deep basis fitting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 71–80, 2020. 2
- [35] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 3
- [36] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 2
- [37] Kourosh Sartipi, Tien Do, Tong Ke, Khiem Vuong, and Stergios I Roumeliotis. Deep depth estimation from visual-inertial slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020. 2
- [38] Nick Schneider, Lukas Schneider, Peter Pinggera, Uwe Franke, Marc Pollefeys, and Christoph Stiller. Semantically guided depth upsampling. In *German conference on pattern recognition*, pages 37–48. Springer, 2016. 18
- [39] Shreyas S Shivakumar, Ty Nguyen, Ian D Miller, Steven W Chen, Vijay Kumar, and Camillo J Taylor. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 13–20. IEEE, 2019. 2, 3, 6, 18
- [40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6, 12, 13, 17, 18
- [41] 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, 6, 12, 13
- [42] 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, 18
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3
- [44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [45] Alex Wong, Safa Cicek, and Stefano Soatto. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):1495–1502, 2021. 2, 3, 4, 5, 6, 7, 8, 14, 16, 17, 18, 19

- [46] Alex Wong, Xiaohan Fei, Byung-Woo Hong, and Stefano Soatto. An adaptive framework for learning unsupervised depth completion. *IEEE Robotics and Automation Letters*, 6(2):3120–3127, 2021. [2](#), [4](#), [6](#), [13](#), [18](#)
- [47] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 2020. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [12](#), [13](#), [14](#), [16](#), [17](#), [18](#), [19](#)
- [48] Alex Wong and Stefano Soatto. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5644–5653, 2019. [13](#)
- [49] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2811–2820, 2019. [2](#), [18](#), [19](#)
- [50] Yanchao Yang, Alex Wong, and Stefano Soatto. Dense depth posterior (ddp) from single image and sparse range. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3353–3362, 2019. [2](#), [3](#), [6](#), [8](#), [14](#), [18](#), [19](#)
- [51] Yinda Zhang and Thomas 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](#)
- [52] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. [16](#)
- [53] Xingxing Zuo, Nathaniel Merrill, Wei Li, Yong Liu, Marc Pollefeys, and Guoquan Huang. Codevio: Visual-inertial odometry with learned optimizable dense depth. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021. [2](#)

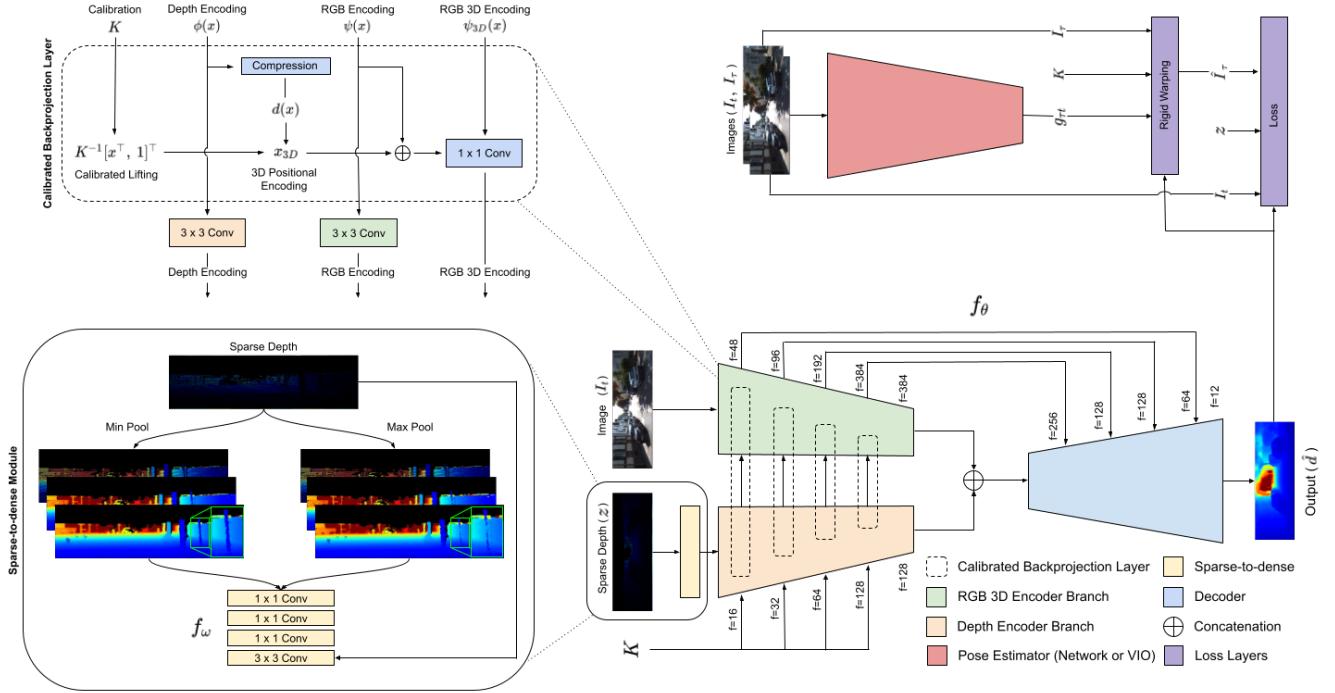


Figure 7: *System diagram during training.* We assume we are given monocular video sequences, synchronized sparse point clouds projected onto the image plane as 2.5D depth maps, and camera calibration. A training sample is therefore (I_t, I_τ, z, K) . Sparse depth inputs (z) are fed to our sparse-to-dense module (f_ω) to yield a dense or quasi-dense representation. Along with image (I_t) and camera calibration matrix (K), it is then fed into our depth completion network (f_θ) comprised of calibrated backprojection layers to produce dense depth prediction \hat{d} . Relative pose ($g_{\tau t}$) between images I_t and I_τ can be estimated from a VIO or a network. In the case of the latter, pose can be jointly learned with depth. We note that pose is only needed to give the reconstruction \hat{I}_τ for constructing the loss function and is not needed during inference.

Code available at: <https://github.com/alexklwong/calibrated-backprojection-network>.

Supplementary Materials

Summary of contents. In Sec. A, we provide an overview of our full system and more details on our loss function. We also provide the kernel sizes used in our sparse-to-dense module, augmentations used during training and our learning rate schedule to reproduce our results on KITTI [41], VOID [47], and NYUv2 [40]. In Sec. B, we visualize and compare features learned by our proposed sparse-to-dense module to those from typical convolutional block, and show that our sparse-to-dense module yields a much denser representation for the depth completion network to ingest. In Sec. C.1, we consider the possibility of miscalibration and examine the sensitivity of our model to changes in intrinsics parameters i.e. *incorrect* calibration. We show that our model is robust to reasonable ranges of calibration error. In Sec. C.2, we study the sensitivity of our model to changes

in sparse depth input density levels and demonstrate that we are robust even when sparse point cover only 0.15% of the image space. In Sec. D, we discuss our method’s ability to generalize to test time sensor platforms with a different camera than the one used in training. Finally, in Sec. E, we show that we can beat several supervised methods on KITTI online leaderboard and that we rank 5th amongst *all methods* for the iMAE metric.

A. System Overview

Fig. 7 shows a diagram of our full system. Our model takes an RGB image I , a sparse depth map z , and the camera intrinsics matrix K as input. First, the sparse depth map z is fed into our sparse-to-dense module f_ω to obtain a dense or quasi-dense representation (Sec. 2.1, main text). Then, the depth representation $f_\omega(z)$, RGB image I , and intrinsics K are fed into the depth completion network f_θ , which is comprised of an encoder with calibrated backpro-

jection layer followed by a decoder (Sec. 2.2, main text). Each calibrated backprojection realizes the backprojection process into 3D camera space by performing calibrated lifting of pixel coordinates using K , and projecting the depth representation to 1 dimension and multiplying it with the lifted coordinates – result of which is a 3D positional encoding of the scene structure.

To yield a unified depth and RGB representation, the 3D positional encoding from the depth branch is passed laterally to the RGB branch to enable association between each RGB feature and its 3D position. By doing so, we introduce 3D structure as an architectural inductive bias, which allows the network to reason about “close” points in the 2D image topology that are actually far in 3D scene topology. The RGB 3D representation is finally fed through the decoder to produce the final depth prediction \hat{d} .

A.1. Loss Function

To train our model, we assume the availability of previous and next RGB frames I_τ of the given image I or I_t (to denote the current time frame) where $\tau \in T \doteq \{t-1, t+1\}$. During training, we estimate the relative pose $g_{\tau t}$ between images at time t and τ . Using I_τ , K and $g_{\tau t}$, we can create the reconstruction \hat{I}_t of I_t via reprojection (Eqn. 4, main text) to enable an unsupervised loss (Eqn. 6-9, main text), which include a photometric reconstruction term, a sparse depth reconstruction term and a local smoothness term.

We note that the photometric term can be replaced with more sophisticated measures of reprojection error [14] and additional regularizers such as pose consistency [47] or adaptive regularization weighting schemes [48, 46] – which would likely boost performance even more. However, we choose a simple loss to demonstrate the efficacy of our novel architecture. We note that $g_{\tau t}$ can be obtained by the means of a visual inertial odometry (VIO) system or a pose network if the VIO is not available. In the case where pose is obtained from network, the pose network can be trained jointly with our depth completion network (KBNet). Relative pose is learned as a byproduct of minimizing Eqn. 6 in main text. Also, since $g_{\tau t}$ is only need for reprojection during training; hence, the VIO system and the pose network are not necessary for inference. Because our network is fast and light-weight (16ms run time per image, 6.9M parameters and 2.6GB memory as benchmarked on 1216×352 images from KITTI [41]), it can be deployed with a VIO system to learn online.

A.2. Implementation and Training Details

We optimized our networks using Adam [21] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained for a total of 60 epochs on KITTI [41], 15 epochs on VOID [47], and 15 epochs on NYUv2 [40]. We use a batch size of 8 with 768×320 crops for KITTI, 640×480 for VOID and 576×416 for NYUv2.

Epochs	Learning Rate
KITTI	
0 to 2	5×10^{-5}
2 to 8	1×10^{-4}
8 to 20	1.5×10^{-4}
20 to 30	1×10^{-4}
30 to 45	5×10^{-5}
45 to 60	2×10^{-5}
VOID	
0 to 10	1×10^{-4}
10 to 15	5×10^{-5}
NYUv2	
0 to 10	1×10^{-4}
10 to 15	5×10^{-5}

Table 6: Learning schedule for KITTI, VOID, and NYUv2.

Dataset	Min Pool	Max Pool
KITTI [41]	5, 7, 9, 11, 13	15, 17
VOID [47]	15, 17	23, 27, 29
NYUv2 [40]	15, 17	23, 27

Table 7: Min pool and max pool kernel sizes for our sparse-to-dense module Kernel sizes for VOID [47] and NYUv2 [40] are larger because the point cloud generated from VIO [11] is much sparser than that of LIDAR in KITTI [41].

For KITTI, we choose $w_{ph} = 1$, $w_{co} = 0.15$, $w_{st} = 0.95$, $w_{sz} = 0.6$, and $w_{sm} = 0.04$; for VOID and NYUv2, we set $w_{sz} = 2$ and $w_{sm} = 2$. Kernel sizes for our sparse-to-dense (S2D) module are shown in Table 7 for each dataset. We detail our learning rate schedule for each dataset in Table 6. For data augmentations on KITTI, we performed random horizontal shifts to the image and depth map and randomly removed between 60% to 70% of the sparse points. For VOID and NYUv2, we randomly removed 30% to 60% of the sparse points. Augmentations are enabled 100% of the time up for VOID and NYUv2. For KITTI it is applied 100% of the time up to the 50th epoch and decreased by half every 5 epoch up to 60 epochs. Each augmentation has a 50% probability of being applied.

B. Features Learned by Sparse-to-Dense

In Sec 2.1 of the main text, we proposed a sparse-to-dense module (S2D) to learn a dense or quasi-dense representation of the sparse depth inputs. S2D utilizes a series of min and max pooling layers of various kernel sizes to densify the sparse depth inputs (for a list of kernel sizes used

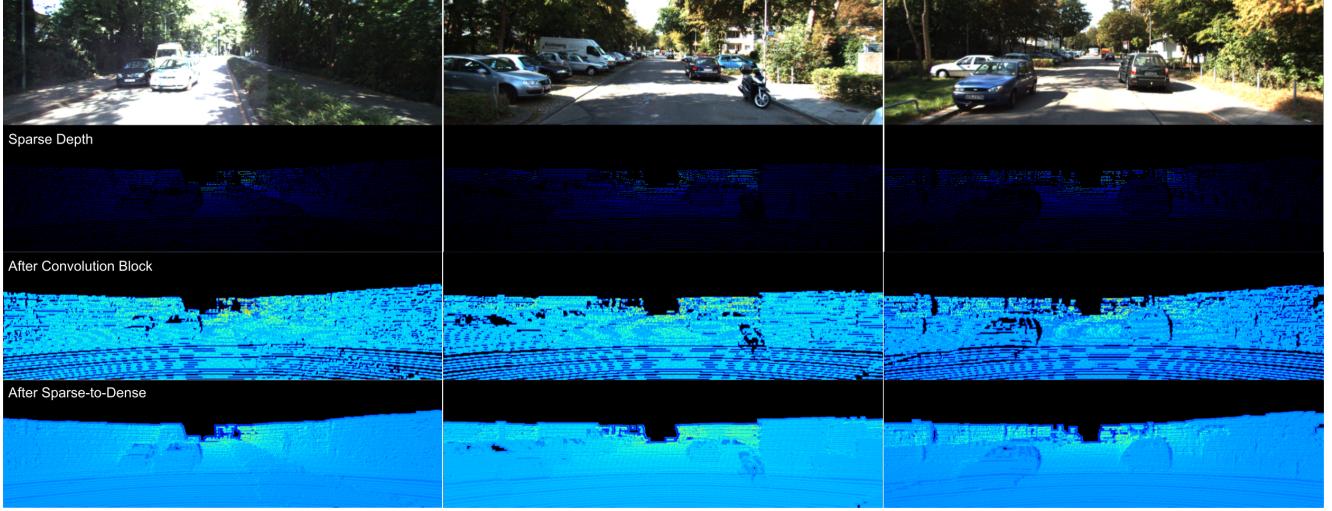


Figure 8: *Visualization of depth features.* Row 3: “After Convolution Block” denotes the depth features produced by a typical first convolutional block used by [26, 47, 50] *without* any form of densification. Row 4: “After Sparse-to-Dense” denotes the depth features learned by the proposed sparse-to-dense (S2D) module. Those learned without our module are still sparse; whereas S2D produces a dense or quasi-dense representation before it reaches the depth completion network. This alleviates the network from having to densify or propagate the sparse signal, making the overall architecture more efficient.

for each dataset, please see Table 7). To balance the trade-off between density and detail (large vs. small kernel sizes), and near and far structures (min vs. max pooling), we concatenate the pooled results and learn three 1×1 convolutions. The output of which is fused with the input sparse depth using a 3×3 convolution to “fill in the gaps”.

Fig. 8 shows visualizations of features learned by S2D and a comparison to the features learned by typical convolutional e.g. ResNet or VGG blocks used by [26, 47, 50]. Row 2 of Fig. 8 shows that despite passing through several convolutional layers ($\approx 10K$ to $20K$ parameters), the representation obtained by a typical convolution block is still sparse; so the later layers will still have many zero-activations and must continue to densify the features. In contrast, using our proposed S2D (≈ 900 parameters), the depth representation learned is dense or quasi-dense before reaching the depth completion network (row 3). This enables non-zero activations in the later layers, which allows the network to use its early convolutions for learning scene geometry rather than densification.

We note that our sparse-to-dense module may bear some resemblance to Spatial Pyramid Pooling (SPP) employed in classification [18] or stereo matching [2]. However, we note that [18] used SPP with max pooling to ensure that feature map sizes are consistent for different input sizes. [2] used average pooling to increase receptive field. Both use cases are intended for dense input. We discussed the drawbacks of max pooling [18] in Sec. 2.1 of the main text and showed in Table 3 of main text that SPP underperforms compare to our S2DM. Also we note that using average pooling [2] will

destroy the signal because the kernel will convolve and average over mostly zeros. The work that is most similar to our S2D module is the SPP for depth completion proposed by [45]. However, [45] only uses max pooling which decimates the detail of nearby structures.

C. Sensitivity Studies

In this section, we provide additional studies to quantify the sensitivity of our model to incorrect calibration and various sparse depth density levels.

C.1. To Incorrect Calibration

We showed in Table 5 of the main text that our method generalizes well when given the *correct* calibration at test time. To consider the scenario of a miscalibrated camera, we studied the sensitivity of our model to *incorrect* calibration on the KITTI dataset (outdoor scenarios) in Fig. 5 in the main text (also here in Fig. 10). Now, we further extend the sensitivity study to the indoor setting by conducting a similar sensitivity study on the VOID dataset (Fig. 11). To this end, we consider changes to the focal length (f) and principal point (c_x, c_y) parameters to create erroneous intrinsic calibration matrices for input to a pretrained model on VOID.

The overall trend for indoor setting, (VOID, Fig. 11) is similar to that of outdoor setting (KITTI, Fig. 10). For both indoors and outdoors, our model is robust to changes in principal point parameters (c_x, c_y) – increasing or decreasing them by up to 25% has little effect on performance.

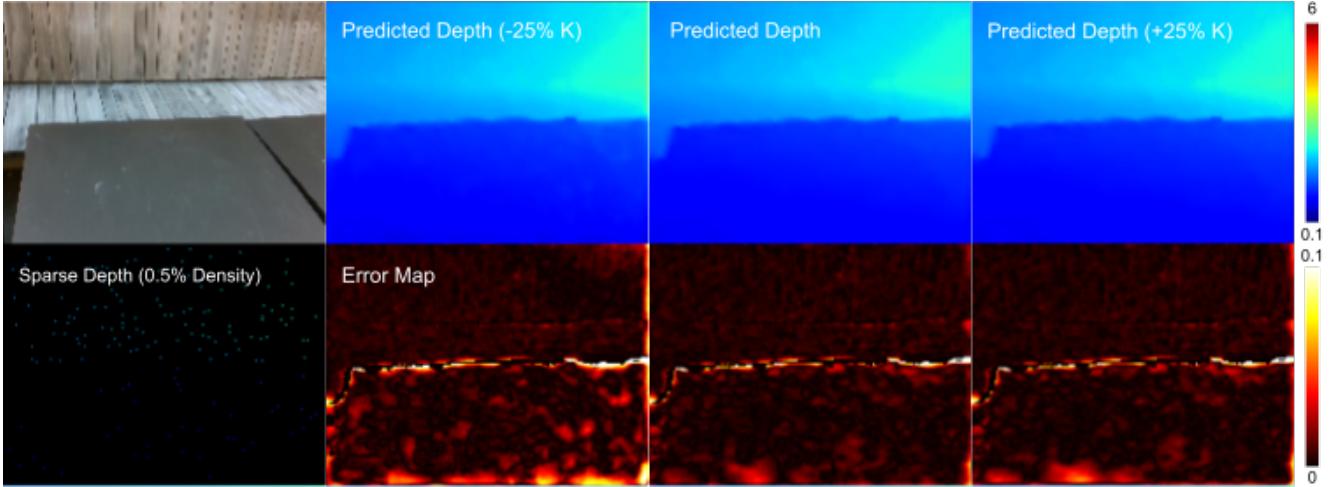


Figure 9: *Visualization of predicted depth for incorrect calibration.* -25% K denotes 25% decrease to intrinsics parameters and +25% K denotes 25% increase. Overall error in -25% is increased (slight brighter shade of red). Larger errors caused by incorrect intrinsics is generally located at the edge of the depth map. +25% have little effect on our predictions. This is because decreasing focal length causes surfaces to be distorted, which in turn affect depth predictions. On the other hand, increasing focal length packs points closer together, which is less detrimental in comparison.

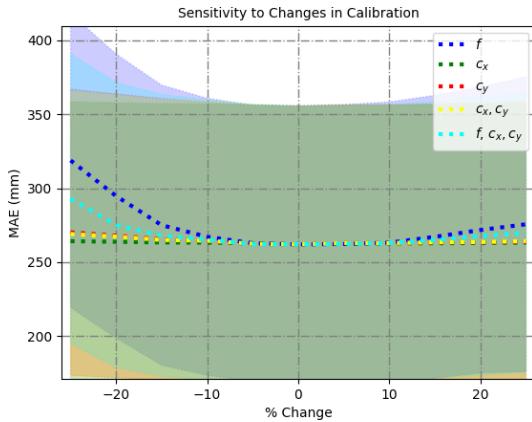


Figure 10: *Sensitivity to changes in calibration on KITTI.* Focal length and principal point are altered to test sensitivity to changes in intrinsics parameters. Our method is robust to change up to $\approx 10\%$ change. After which, performance degrades. We note that changes in principal point (c_x, c_y) have little effect; whereas decreasing focal length (f) causes large drop in performance.

This is because these parameters shifts the optical center so they do not affect the overall structure of the scene. We note that for large values outside of reasonable perturbation range will cause the performance to decrease.

Unlike its behavior with changes in the principal point, the model degrades when focal length (f) is decreased. For both indoors and outdoors, we are robust up to 10% decrease in focal length, after which error will increase. We

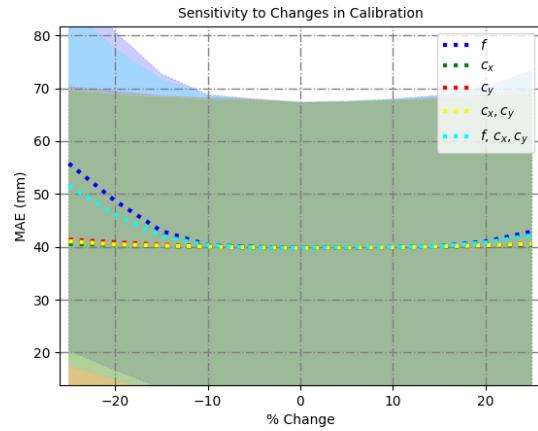


Figure 11: *Sensitivity to changes in calibration on VOID.* Focal length and principal point are altered to test sensitivity to changes in intrinsics parameters. Our method is robust to change up to $\approx 10\%$ change. After which, performance degrades. We note that changes in principal point (c_x, c_y) have little effect; whereas decreasing focal length (f) causes large drop in performance.

note that the performance drop is asymmetric, our model is robust to increases in focal length up to 20%. The reason for this phenomenon is as follows: Geometrically, decreases in focal length will cause points to backproject to a wider field of view, which distorts surfaces by pushing points that belong to the same surface far from each other. On the other hand, increases in focal length will cause points to pack tighter together. This does not disrupt the scene struc-

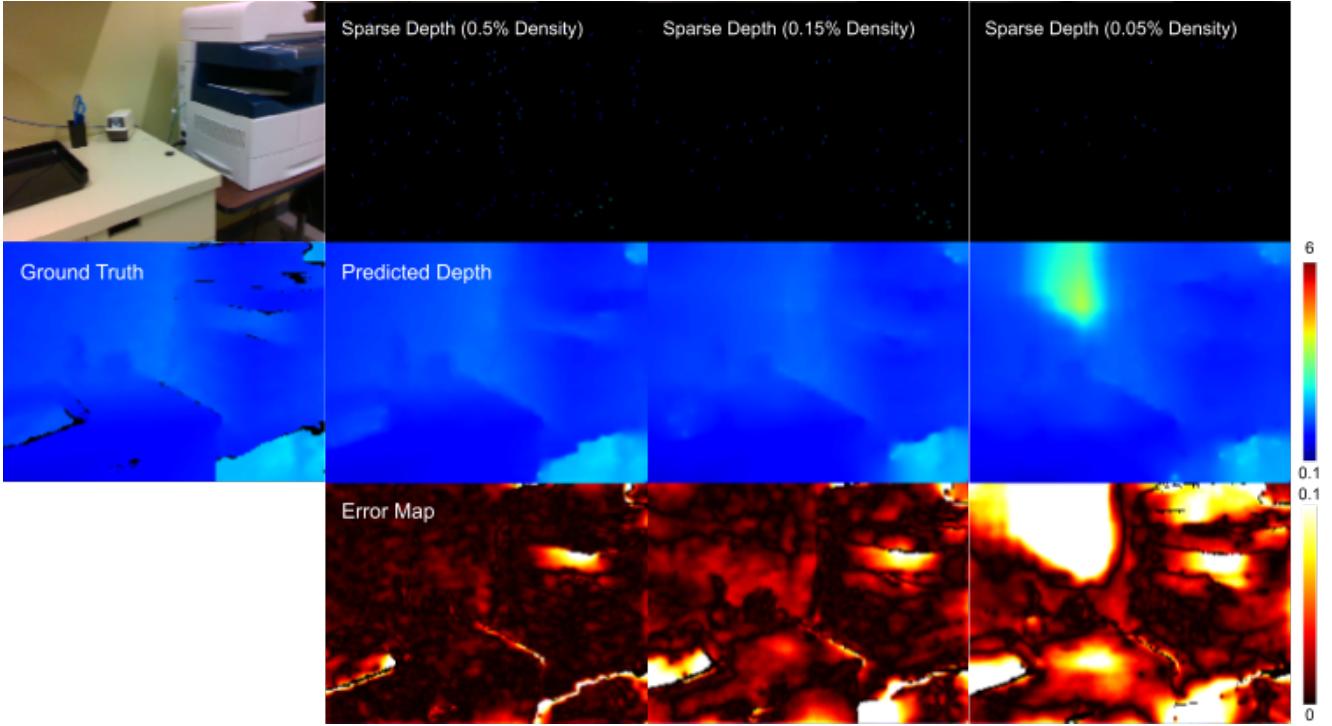


Figure 12: *Visualization of predicted depth for various density levels on VOID.* Columns 1, 2: Our method works well for density levels of 0.5% and 0.15%. Column 3: The quality of predicted depth begins to degrade in far homogeneous regions where there are no sparse points e.g. wall when density level drops to 0.05%.

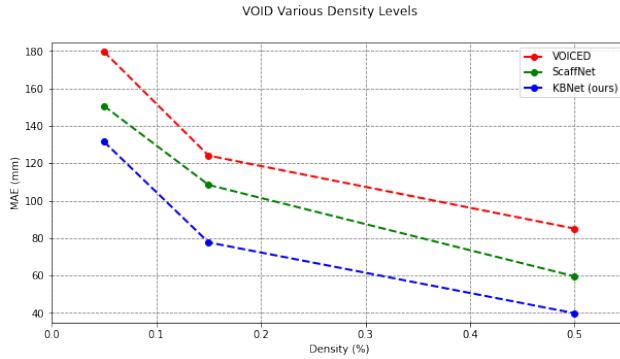


Figure 13: *VOID test set across various density levels.* We compare our method against VOICED [47] and ScaffNet [45] on the VOID test set for various density levels (0.5%, 0.15%, 0.05%). In terms of MAE, our method performs better than other methods across all density levels.

ture for small values, but for large values, points will get squashed together; this is demonstrated by the small uptick in error when increasing focal length by 20 to 25%.

We note that these values are well out of the typical range of calibration error and should not be of concern. For example when using off-the-shelf calibration packages that implements [52] to calibrate our camera, we obtained a standard error of $\approx 0.6\%$, which yields $\pm \approx 1.1\%$ margin of

Method	MAE	RMSE	iMAE	iRMSE
0.50% Density				
VOICED [47]	85.05	169.79	48.92	104.02
ScaffNet [45]	59.53	119.14	35.72	68.36
Ours	39.80	95.86	21.16	49.72
0.15% Density				
VOICED [47]	124.11	217.43	66.95	121.23
ScaffNet [45]	108.44	195.82	57.52	103.33
Ours	77.70	172.49	38.87	85.59
0.05% Density				
VOICED [47]	179.66	281.09	95.27	151.66
ScaffNet [45]	150.65	255.08	80.79	133.33
Ours	131.54	263.54	66.84	128.29

Table 8: *Sensitivity study on various sparse depth density levels on VOID.* We train a single model on VOID using sparse depth maps of 0.50% density and evaluate it on 0.50%, 0.15%, 0.05% density test sets. As expected, performance degrades as the input becomes more sparse. Overall, we perform better than [45, 47]; however, at 0.05%, [45] performs better on the RMSE metric.

error for a 95% confidence interval. Nonetheless, there exists the risk of using the wrong calibration; however, we

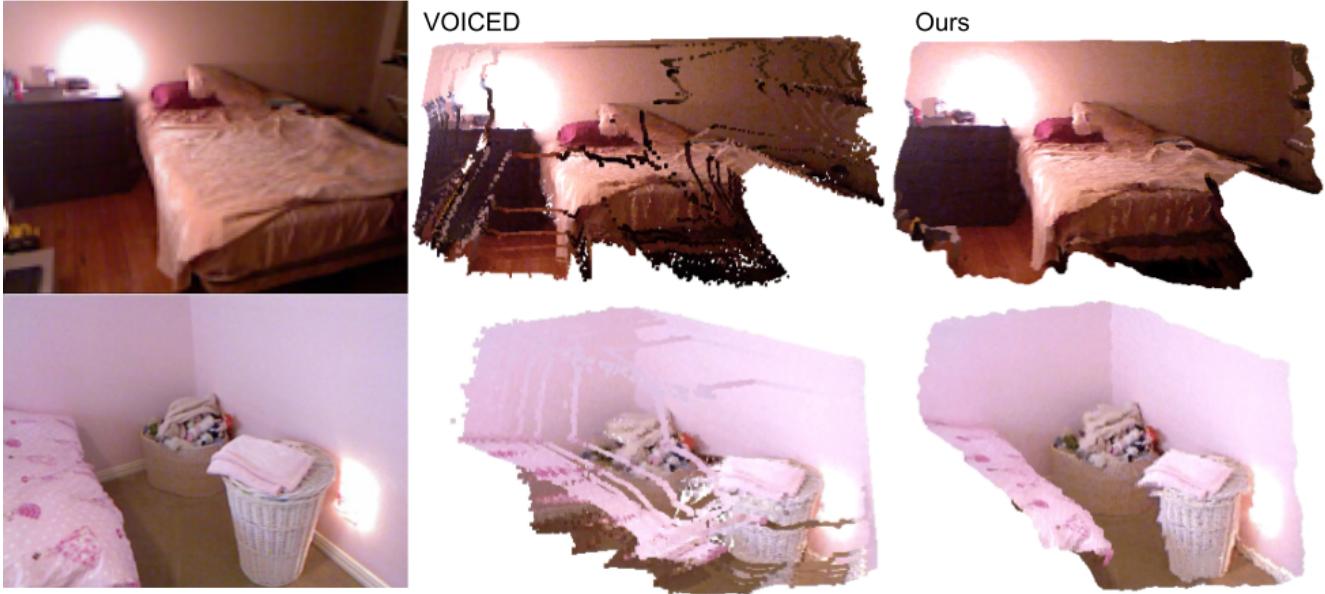


Figure 14: *Qualitative results on generalization to novel scenes captured by a different sensor platform.* We trained our model on VOID [47] (captured by Intel RealSense) and tested the model on NYUv2 [40] (captured by Microsoft Kinect). We also used a pretrained model (on VOID) of [47] as the baseline and tested it on NYUv2. Here, we show the predicted depth as point clouds, backprojected to 3D and colored. [47] predicted a distorted scene where the points are bowed towards the camera; whereas, while our predictions are not perfect, they are reasonable.

believe this trade-off is well worth the performance boost provided by the proposed architecture.

Fig. 9 shows a visualization of depth predicted by our model when using erroneous calibration. -25% K denotes a 25% decrease to focal length and principal point and +25% K denotes a 25% increase to both. As we can see, the larger errors are typically located along the border of the predicted depth map; there is also a slight increase in error (brighter shade of red) for the entire scene. Increasing intrinsics by 25% affects the output less significantly, but nonetheless we observe an increase in errors.

C.2. To Various Density Levels

In Table 8, we consider three different levels of density for the sparse depth inputs, 0.50%, 0.15%, 0.05% of the image space, that are provided by the VOID dataset [47]. To this end, we train a *single* model on VOID using sparse depth maps of 0.50% density and evaluate it on 0.50%, 0.15%, 0.05% density test sets. We also compare our method against [45, 47] under these density levels.

As expected, as density decreases, our performance also degrades. However, we still outperform both [45, 47] under all three levels, see Fig. 13. We note that at the sparsest setting of 0.05%, [45] does beat us on the RMSE metric. The reason for this is that we selected the kernel sizes for our model based on the sparsity level of 0.5%; therefore, when testing it on 10× sparser point cloud, our depth representa-

tion will be more sparse as well, which limits the potential of our calibrated backprojection layers. In contrast, [45] proposed a network to first estimate the dense coarse topology. This phenomenon is also observed in KITTI, shown in Table 3 of the main text, where we removed our sparse-to-dense module and we observed a significant drop in performance.

Fig. 12 shows qualitative evaluations on the three density levels. For 0.50%, error is low overall and the shape of the recovered scene resembles that of the ground truth. When we decrease density to 0.15%, we observe slight blurring in object shapes and increased errors in homogeneous regions. At 0.05%, we begin to observe artifact such as the green “blob” corresponding to the wall with more exaggerated errors in homogeneous regions. This is because locally the textureless surfaces give little to no information on object shape. Without sparse depth to anchor their values, they can be arbitrary. In this case the “empty” region is predicted as far.

D. Generalization to Other Sensor Platforms

In Sec. 3.4 of the main text, we discussed our ability to generalize to other sensor platforms that may use a different test time camera than one used to collect training data. In Table 5 of the main text, we showed quantitatively that we generalize better than the baseline. Here, we demonstrate this qualitatively in Fig. 14.

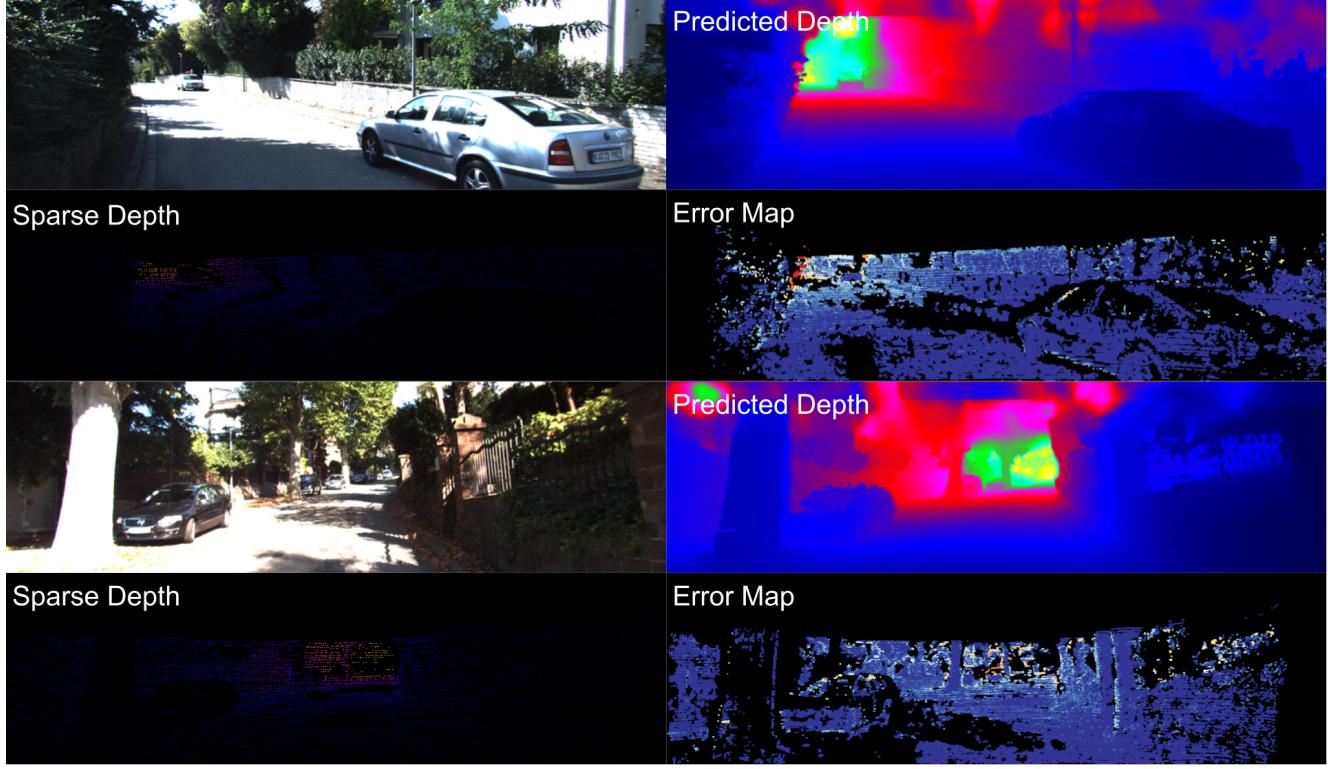


Figure 15: Qualitative results on KITTI depth completion benchmark.

Method	MAE	RMSE	iMAE	iRMSE
ADNN [6]	439.48	1325.37	3.19	59.39
Morph-Net [7]	310.49	1045.45	1.57	3.84
CSPN [5]	279.46	1019.64	1.15	2.93
KBNet (Ours)	256.76	<i>1069.47</i>	<i>1.02</i>	2.95
SS-S2D [26]	249.95	814.73	1.21	2.80
DeepLiDAR [32]	226.50	758.38	1.15	2.56
PwP [49]	235.73	785.57	1.07	2.52
UberATG-FuseNet [3]	221.19	752.88	1.14	2.34
RGB_guide&certainty [42]	215.02	772.87	0.93	2.19
DDP [50]	203.96	832.94	0.85	2.10
CSPN++ [4]	209.28	743.69	0.90	2.07
NLSPN [30]	199.59	741.68	0.84	1.99

Table 9: KITTI supervised depth completion benchmark. Results are directly taken from online leaderboard. Note: SS-S2D [26] and DDP [50] compete in both supervised and unsupervised benchmarks. Our results are italicized. Despite being an unsupervised method, our method beats some supervised methods [6, 7] and our iMAE score (1.02) is ranked 5th amongst supervised methods.

To this end, we trained our model on VOID [47] (captured by Intel RealSense) and tested the model on NYUv2

Method	MAE	RMSE	iMAE	iRMSE
SGDU [38]	605.47	2312.57	2.05	7.38
SS-S2D [26]	350.32	1299.85	1.57	4.07
IP-Basic [22]	302.60	1288.46	1.29	3.78
DFuseNet [39]	429.93	1206.66	1.79	3.62
DDP* [50]	343.46	1263.19	1.32	3.58
VOICED [47]	299.41	1169.97	1.20	3.56
AdaFrame [46]	291.62	1125.67	1.16	3.32
SynthProj* [25]	280.42	1095.26	1.19	3.53
ScaffNet* [45]	280.76	1121.93	1.15	3.30
KBNet (Ours)	<i>256.76</i>	<i>1069.47</i>	<i>1.02</i>	2.95

Table 10: KITTI unsupervised depth completion benchmark. Results are directly taken from online leaderboard. Note: SS-S2D [26] and DDP [50] compete in both supervised and unsupervised benchmarks. Our method outperforms is trained only on KITTI, but still the state of the art [45] (trained on KITTI and Virtual KITTI [13]) by an average of 8.8% across all metrics. * denotes methods that use additional synthetic data for training.

[40] (captured by Microsoft Kinect). We similarly trained the baseline [47] on VOID and tested it on NYUv2. Fig. 14 shows the predicted depth, backprojected to the point clouds

in 3D and colored. As we can see, [47] predicted a distorted scene; in contrast, ours is not perfect, but reasonable. This demonstrates the benefit of taking calibration as input. It allows the model to generalize well when it is deployed to a sensor platform where the camera that is used is *different* than the one used for training. We also note that neither models have been trained on NYUv2 which features a different scene distribution than that of VOID.

E. KITTI Depth Completion Benchmark

In Sec. 3.3 of the main text, we compare our method against *unsupervised* methods on the KITTI online leaderboard. Here, we show quantitative comparisons against both supervised (Table 9) and unsupervised (Table 10) methods. Results and method names are directly taken from the KITTI online leaderboard. Here we refer to our method as KNBet, as listed on the leaderboard. We note that SS-S2D [26] and DDP [50] compete in both supervised and unsupervised benchmarks. Additionally, we provide high resolution examples of our output in Fig. 15.

Despite being trained without ground-truth annotations, Table 9 shows that our method is competitive even amongst supervised method. We outperform some supervised methods [5, 6, 7] across most metrics. We note that our method significantly improves the iMAE and iRMSE metrics, to the point where we are comparable to some of the supervised methods for close range performance. Our iMAE score, which penalizes mean error in close range regions, is ranked 5th overall amongst both supervised and unsupervised methods. To the best of our knowledge, we are the first work in unsupervised depth completion to demonstrate comparable performance to supervised methods. We note that supervised methods are generally more computationally expensive with high model complexity e.g. in terms of number of parameters, [30] uses 25.84M, [32] 53.4M, and [49] 28.99M; whereas we only use 6.9M.

Compared to unsupervised methods (Table 10), we rank first amongst all methods with the best scores across all metrics. Our model even beat methods [25, 50, 45] that use additional synthetic data (Virtual KITTI [13]) for training, amongst which is the state of the art [45]. Despite this, we beat [45] by an average of 8.8% across all metrics while using 11.5% fewer parameters. These results demonstrates the potential of our method to bridge the gap between supervised and unsupervised method. Moreover, our network is light-weight and can be deployed on VIO system [11]. While there is still a long road ahead, these results show a lot of promise in enabling unsupervised methods to learn online and to be used for real-time application for low-cost hardware systems.