Review

A Comprehensive Survey of Depth Completion Approaches

Muhammad Ahmed Ullah Khan[1,2,3,†], Danish Nazir[1,2,3,†], Alain Pagani[3], Hamam Mokayed [4], Marcus Liwicki[4], Didier Stricker [1,3] and Muhammad Zeshan Afzal [1,2,3*]

[1] Department of Computer Science, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany; muhammad_ahmed_ullah.khan@dfki.de(A.U.K.); danish.nazir@dfki.de (D.N.); muhammad_zeshan.afzal@dfki.de (M.Z.A.); didier.stricker@dfki.de (D.S.);
[2] Mindgarage, Technical University of Kaiserslautern, 67663 Kaiserslautern, Germany
[3] German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany, alain.pagani@dfki.de (A.P.);
[4] Department of Computer Science, Luleå University of Technology, 971 87 Luleå, Sweden; hamam.mokayed@ltu.se (H.M.); marcus.liwicki@ltu.se (M.L.);
* Correspondence: muhammad_zeshan.afzal@dfki.de
† These authors contributed equally to this work.

Abstract: Depth maps produced by LiDAR based approaches are sparse. Even high-end LiDAR sensors produce highly sparse depth maps, which are also noisy around the object boundaries. Depth completion is the task of generating a dense depth map from a sparse depth map. While the traditional approaches focus on directly completing this sparsity from the sparse depth maps, modern techniques use RGB images as a guidance tool to resolve this problem. Whilst many others rely on affinity matrices for depth completion. Based on these approaches, we have sub-divided the literature into two major categories; traditional approaches and backbone-based approaches. The latter is further sub-divided into two-branch, and spatial propagation approaches. The two-branch approaches still have a sub-category named guided-kernel approaches. In this paper, for the first time ever we present a comprehensive survey of depth completion methods. We present a novel taxonomy of depth completion approaches, review and detail different state-of-the art techniques within each category for depth completion of LiDAR data, and provide quantitative results for the approaches on KITTI and NYUv2 depth completion benchmark datasets.

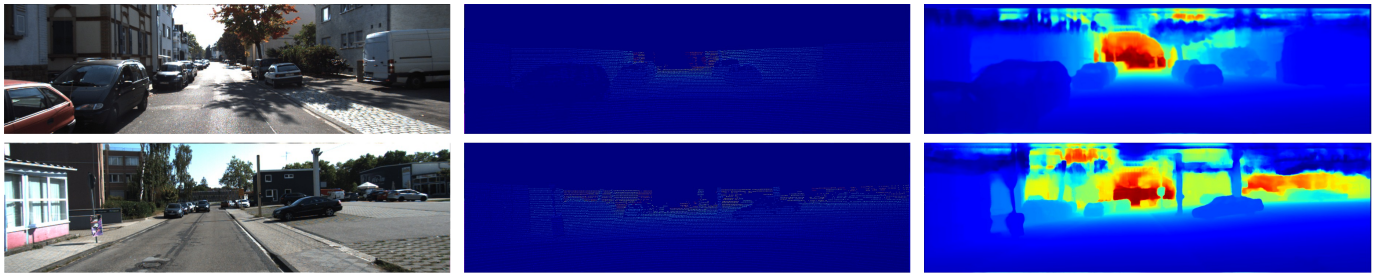
Keywords: Depth Completion; Depth Maps; Image-Guidance

1. Introduction

Depth maps are critical to a variety of computer vision applications such as autonomous driving [1–3], robot navigation [4,5], augmented reality [6–8],virtual reality [9]. Tasks like object detection, obstacle avoidance [10], 3D scene reconstruction [11–13] require dense depth maps for accurate prediction. Various depth sensors like depth cameras, 3D LiDAR and stereo cameras capture the depth information. Among these, LiDAR sensors provide the most accurate depth information. However, the depth maps generated by these devices are sparsely distributed (1) compared to a medium resolution RGB image (about 5% density [14]). Also, current LiDAR sensors obtain measurements at only 64 scan lines in the vertical direction. This sparsity significantly impacts the performance of LiDAR based applications. Predicting dense depth maps from these sparse ones is critical for both the industry and academia.

To resolve the problem of depth completion, many different approaches have been developed. Traditional approaches [15–17] concentrate on retrieving dense depth maps from the sparse ones without the guidance of an image. Uhrig et al. [18] propose a sparsity invariant CNN to deal with sparse data or features. Eldesokey et al. [19] solve depth completion by generating a full depth as well as a confidence map with normalized convolution. But, these approaches are limited and lose depth details and semantic information without the availability of multi-modal data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1



(a) RGB Image **(b)** LiDAR sparse Depth Map **(c)** Prediction

Figure 1. First Column shows the RGB images from two different scenes, the middle column contains the sparse depth maps produced from LiDAR. Last column shows the predicted dense depth maps for the corresponding scenes.

Image guided methods show significant improvement in results compared to the conventional depth-only techniques. Qiu et al. [20] train a network to predict surface normal using the color image and depth map and further use the recovered surface normal to guide depth completion. CSPN [21] refine coarse depth maps with spatial propagation network using affinity matrices at the end of its Unet [22]. CSPN++ [23] additionally improves by learning adaptive convolution kernel sizes and the number of iterations for propagation. However, most of these techniques consider the task as one-stage learning and use naıve fusion approaches resulting in blurred depth maps with unclear boundaries.

Some works construct a two-branch architecture for handling image and depth modalities and then perform fusion like FusionNet [24] and DeepLiDAR [20]. FusionNet extracts local and global features using its two-branch architecture. While, DeepLiDAR takes multi-modal inputs and performs fusion at a multi-scale level, achieving better depth completion results. But both these methods require extra datasets to pretrain their networks.

The content of this paper is organized as follows: Section 2 provides an overview of the fusion strategies and approaches used in the field of depth completion. Section 3 discusses the common indoor and outdoor dataset used for depth completion. Section 4 introduces the metrics used in the field of depth completion and Section 5 presents the state-of-the-art methods in each category. Finally, Section 6 provides the conclusion of this paper.

2. Methodologies

In this section, we will discuss both the approaches to dense depth completion and multi-modal fusion strategies to fuse the multi-modal (RGB, LiDAR, Semantic maps, Surface normals) information. Figure 2 shows the approaches to depth completion. Roughly, the approaches can be divided into two different categories; (1) Traditional Approaches, which utilize only LiDAR sparse depth maps for dense depth completion, and (2) Image-guided Methods, which employ guidance images (RGB, semantic maps, surface normals) to guide the process of depth completion. Image-guided methods are more successful than traditional approaches. However, image-guided methods require the employment of fusion strategies to adaptively fuse the information between different modalities. Therefore, we also discuss multi-modal fusion strategies in this section.

2.1. Traditional Approaches

Traditional approaches can be further classified into single-branch approaches, since they utilize only one branch to process sparse LiDAR data. Earlier approaches [16,18,25] based on convolutional neural networks (CNN) utilized only sparse depth maps to generate dense depth maps. To fill the missing values at invalid regions of sparse depth maps, many hand-crafted features, kernels, interpolation methods [26–30] were introduced. However, the structural information of the scene is lost because of the discontinuity in the depth values. To counter the sparsity of data in sparse depth

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

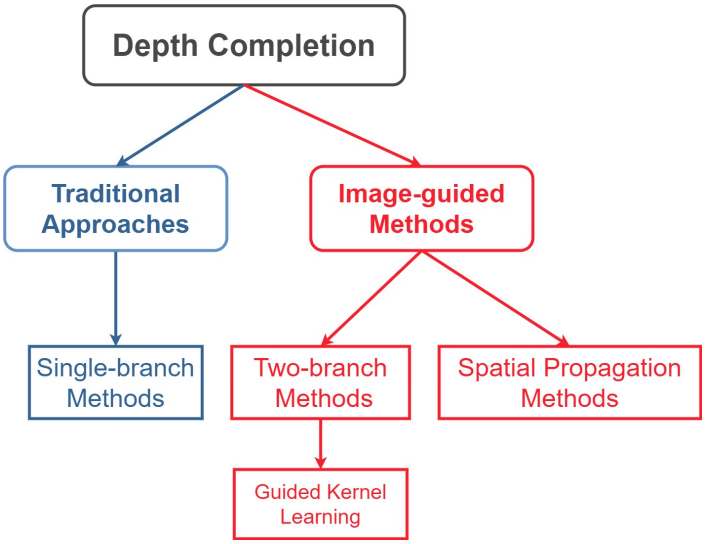


Figure 2. Approaches to Depth Completion problem. Traditional approaches include single-branch (utilize only LiDAR) methods to depth completion. The image-guided methods (two-branch and spatial propagation methods) utilize guidance images (RGB, semantic maps, surface normals) to guide the process of depth completion. The two-branch methods can be further divided into guided kernel learning methods, which aim to learn useful kernels from one modality and apply it to other modalities.

maps, Depth-Net [31] performed nearest-neighbor interpolation in the sparse maps to fill out the holes. Later on, uncertainty-aware CNN's [32] proposed probabilistic normalized convolutions to model the uncertainty in the sparse depth maps. The obvious drawback of these approaches is that without color or semantic image guidance, the predicted depth maps lack clear object boundaries, making them unsuitable for real-time applications.

2.2. Multi-modal Fusion

Multi-modal fusion refers to the approaches and methodologies of fusing sensor information from two or more different sensors to enhance the understanding of the environment. In the context of depth completion, it refers to the process of utilizing information from different modalities including RGB cameras [33,34], surface normal's [20], semantic maps [35,36] etc., to guide the process of dense depth completion. The goal of multi-modal fusion is to leverage different modalities or their feature representations to produce reliable information on the sparse regions of LiDAR depth maps. Following sections cover common fusion techniques for depth completion.

2.2.1. Early Fusion

The idea of early fusion is to integrate the separate raw modalities e.g., RGB camera and LiDAR sensor, into a single unified representation [37]. There exist many methods to compute the unified representation. However, most common methods include point pixel projection between RGB image and LiDAR sparse depth map [38], concatenation or addition of RGB and LiDAR sparse depth map [33,39], etc. The unified representation is then sent to the AutoEncoder for dense depth completion. The pipeline of early fusion is depicted in Figure 3.

2.2.2. Sequential Fusion

Sequential fusion is an extension of early fusion. In the first step, it predicts a dense color depth through an RGB branch consisting of an RGB-only deep neural network. The color depth is a very noisy estimate of dense depth, but it contains the depth information around the object boundaries, e.g., cars and trees, which is missing in LiDAR sparse

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

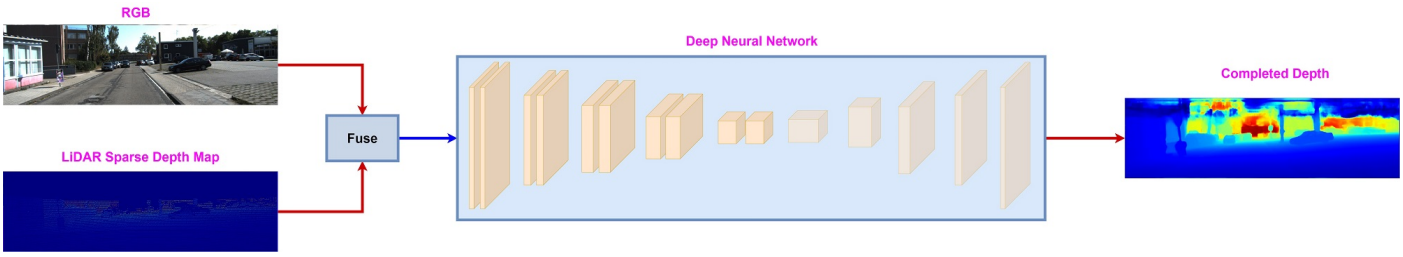


Figure 3. Early fusion between RGB image and LiDAR sparse depth. At first, both modalities are fused and then sent to the Deep Neural Network for dense depth completion.

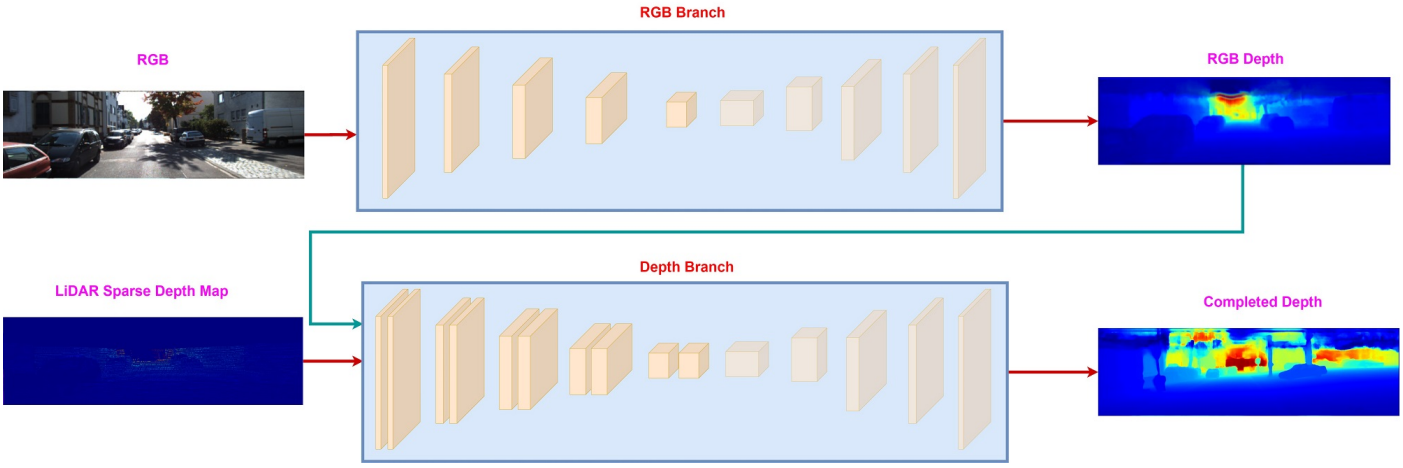


Figure 4. Sequential fusion between RGB image and LiDAR sparse depth map. The RGB branch produces color depth, which along with LiDAR sparse depth map, is sent to depth branch to estimate the final dense depth map.

depth map [33,36,39]. The color depth and LiDAR sparse depth map are sent to the depth branch, which produces the final estimate of the dense depth map. Figure 4 shows the process of sequential fusion between RGB image and LiDAR sparse depth map.

2.2.3. Late Fusion

Unlike early and sequential fusion, the late fusion process both modalities, i.e., RGB color images and LiDAR sparse depth map, independently and fuse them at the final stage. The RGB and depth branches consist of RGB-only and depth-only deep neural networks. The RGB branch outputs a dense depth map focused on color information, whereas the depth branch produces a dense depth map relying more on the LiDAR sparse depth map features [33,36]. The dense depth maps produced by RGB and depth branches are fused to produce the final dense depth map. The final dense depth map combines the strength of both the RGB camera and LiDAR sensor into a single dense depth map. Figure 5 depicts the pipeline of the late fusion for the RGB camera and LiDAR sparse depth map.

2.2.4. Deep Fusion

Deep fusion refers to the fusion performed at the level of the deep neural network. Figure 6 shows the pipeline of the deep fusion between LiDAR sparse depth map and RGB image modalities. Similar to late fusion, the pipeline of deep fusion consists of two separate branches for RGB and LiDAR sparse depth modalities. However, in deep fusion, instead of applying the fusion at the intermediate output of the two branches, it is performed at the feature level throughout the two branches. The fusion follows the decoder-encoder strategy since the features from the RGB decoder are fused at the encoder of the depth branch at multiple stages. Deep fusion only fuses the decoder

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

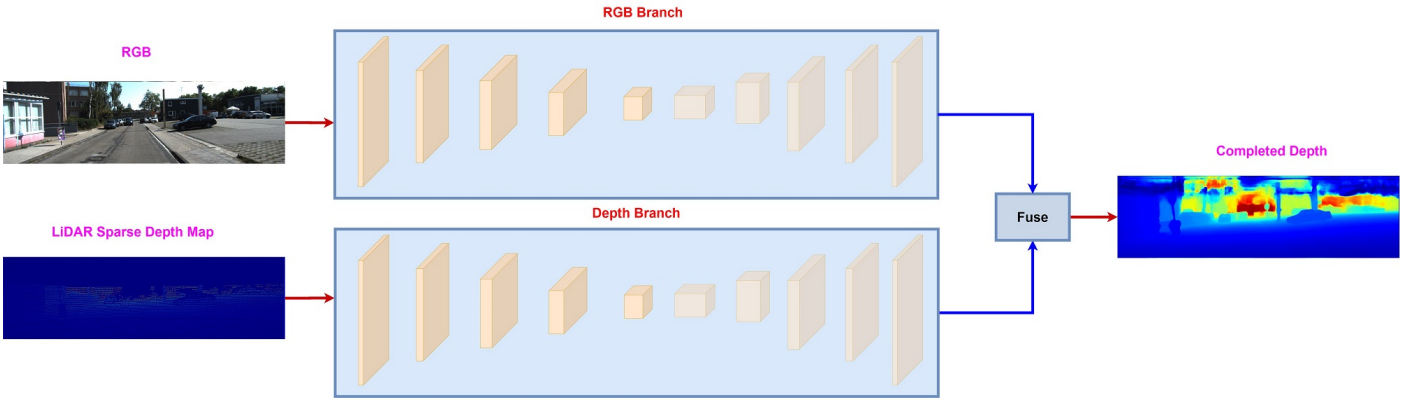


Figure 5. Late fusion between RGB image and LiDAR sparse depth map. It consists of two separate branches to process RGB images and LiDAR sparse depth maps. Both of the branches produce dense depth maps, which are fused to produce a final dense depth map.

features of one modality to another because the decoder contains high-level information, which is used to guide the other modality during dense depth prediction [33,40].

2.3. Image-guided Methods

Image-guided techniques refer to the ones that employ guidance images such as RGB images [33,34], semantic maps [35,36], surface normals [20] and sparse depth map modalities [18] to guide the process of depth completion. These techniques have shown much comprehensive results compared to the traditional approaches.

2.3.1. Two-branch Networks

Two-branch methods refer to the ones that employ two branches for handling the multi-modal information, including RGB images, surface normals, semantic maps and LiDAR sparse depth maps. Each branch treats a single modality separately and then the information from the different branches is fused through multi-modal fusion techniques explained in Section 2.2.

Van Gansbeke, Wouter, et al. [24] propose a two-branch network to extract both the global and local information to produce accurate and comprehensive depth maps. They employ a fusion method based on color image guidance to better incorporate the object information, which significantly improves accuracy. Moreover, confidence masks are learned for both the local and global branches in an unsupervised manner. These masks are then used to weight the depth maps to correct the uncertainty in the depth predictions from both modalities.

DeepLiDAR [20] presents an end-to-end deep learning architecture for accurate image guided depth completion for outdoor scenes using estimated surface normals [41] as intermediate representations to enforce geometric constraints. The sparse depth and image modalities are effectively fused together by the proposed modified two-branch encoder-decoder network [22]. To resolve the issues specific to outdoor scenes, the network predicts a confidence mask to handle artefacts in mixed LiDAR signals near foreground boundaries due to occlusion. Also, it learns attention maps to combine estimates from the color image and surface normal to improve the depth accuracy, especially for distant areas.

Similar to DeepLiDAR [20], to resolve the issues in handling sensor noise and 3D geometric constraints, Xu et al. [42] propose a unified two-stage CNN framework. Firstly, the framework models the geometric constraints between depth and surface normal [41] in a diffusion module. Secondly, similar to [24] it predicts the confidence of sparse LiDAR measurements to reduce the propagation of information due to noise. The surface normals, coarse depth and confidence of LiDAR inputs, predicted by the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

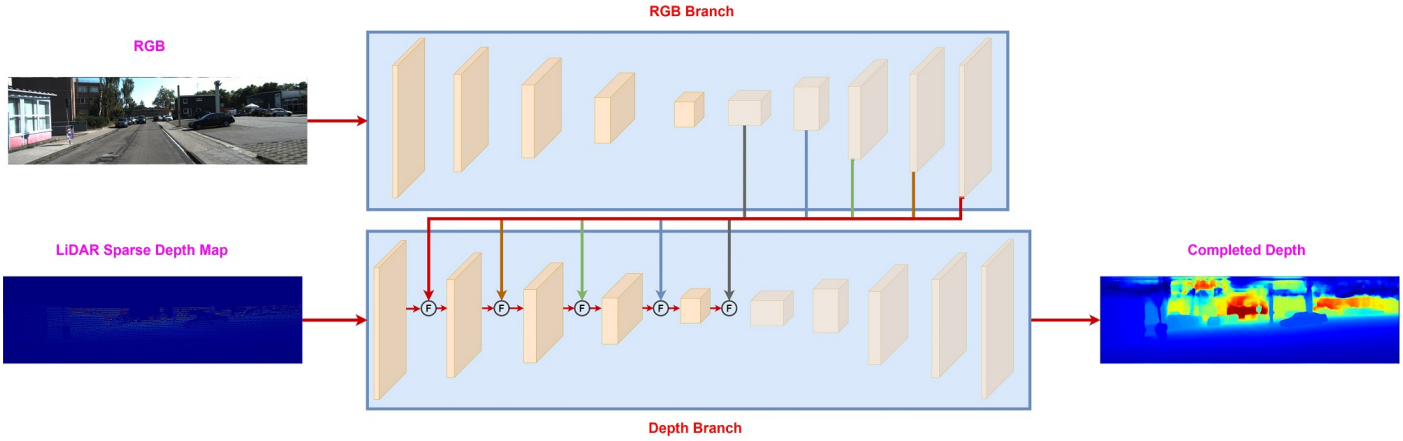


Figure 6. Deep Fusion between RGB image and LiDAR sparse depth map. Each modality is passed from a dedicated branch. The features from decoder of RGB branch are fused into the encoder of the depth branch. The symbol "F" represents the fusion operation. Common choices for fusion operation includes addition or concatenation. However, complex fusion schemes can also be employed. By the guidance of RGB branch, the depth branch produces a final dense depth map.

encoder-decoder backbone [22] are passed to a diffusion refinement module to obtain the final depth completion results.

Unlike the conventional approaches, which make a point estimate, Yang et al. [43] present a deep learning system to estimate the posterior distribution of a dense depth map linked with an image by utilizing sparse range measurements of a LiDAR depth map. Relations between seen images and corresponding depth maps are used to complete the map to get a probability over depth for each pixel in the image. A Conditional Prior Network then combines it with a likelihood term that uses the sparse measurements.

Ma et al. [15] design a deep regression model to directly learn a mapping from a sparse depth map and color image (if available) to a dense depth map. Additionally, they explore a self-supervised training framework for depth completion tasks. The framework requires only the sparse depth and color image sequences, removing the need for dense depth labels during training. This approach also performs better than some of the semi-dense annotation methods.

The standard convolutions fail to model the observed spatial contexts due to sparsity in depth maps. To fully capture the observed spatial contexts, Zhao et al. [44] propose graph propagations. Firstly, they construct multiple graphs at different scales from observed pixels. Then an attention mechanism is applied to the propagation, which allows modeling of the contextual information adaptively. These graph propagations are applied to the depth and image modalities to extract the respective representations. A symmetric gated fusion strategy is proposed to effectively exploit the extracted multi-modal features by learning adaptive gating weights to preserve the original information for one modality and absorb complementary information from the other modality.

Li, Ang, et al. [45] propose a multi-scale guided cascade hourglass network [46] to handle diverse patterns in depth maps efficiently. Unlike the traditional fully convolutional techniques, specialized hourglasses in the cascade network take inputs at different resolutions to predict depth structures at particular scales. An encoder extracts multi-scale features from colour images to provide guidance information for specific structures for all the hourglasses stack. This multi-scale training strategy activates the effect of cascade stages. Also, the division into sub-modules allows replacing the redundant network with a combination of simple architectures.

DenseLiDAR [47] propose a novel real-time pseudo-depth guided depth completion backbone based neural network. They argue that a dense reference depth map is essential to produce accurate dense predictions. The pseudo-depth map is obtained from simple

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

morphological operations and is used to guide the network on three fronts. Firstly, it predicts a residual structure for the output, making it more stable and accurate. Secondly, to rectify the sparse input data and lastly, to enforce a 3D dense structural loss for training the network. Additionally, two new metrics; $RMSE_{GT+}$ and $RMSE_{Edge}$ are proposed for better evaluation of the predicted dense depth maps. The former computes the depth error on a carefully complemented ground truth, while the latter evaluates the accuracy on edge areas of the depth map.

Most of the earlier mentioned image guided depth completion methods use simple concatenation and element wise addition to handle multi-modal fusion. The deep convolutional encoder-decoder architecture [22] designed by Lee et al. [48] incorporates a cross-guidance module for multi-modal feature fusion to overcome the lack of representation power. The two encoders share the information by exchanging the outputs with the guidance module of the other encoder, which applies an attention mechanism to fuse the features. Also, a residual atrous spatial pyramid block (RASP) is proposed to extract highly significant features. This block applies multiple dilated convolutions [49] with different dilation rates in parallel.

Similar to Sparsity Invariant Convolution (SI-Conv) proposed by Uhrig et al. [18] for depth-only completion tasks, Yan et al. [50] propose a novel fusion scheme to effectively fuse the data from image and depth modalities by exploiting the property of image guided depth completion task and data. The technique employs three mask aware operations to process, downscale and fuse the sparse features, where each explicitly considers the distribution of the data and the observation mask of the corresponding feature map. The presented deep neural network processes the two modalities independently, followed by a spatial pyramid fusion block to fuse the features under various receptive fields.

Different to previously discussed approaches which use a typical Convolution Neural Network (CNN) layer, the approach in [19] introduces a novel normalized convolutional layer with a much smaller number of parameters for unguided scenes depth completion on the highly sparse input depth map. It further presents novel methodologies to compute and propagate convolutional confidences to consequent CNN layers. A new loss function is also proposed, minimizing the data error while maximizing the output confidence. The authors also explore several fusion techniques to combine the multi-modal data and integrate structural information in the proposed framework. Additionally, unlike [15] the output confidence is used as auxiliary information to improve the results.

Encouraged by the current approaches in depth completion, which focus on dense guidance, Schuster, René, et al. [51] propose a Sparse Spatial Guided Propagation (SSGP), which is the combination of spatially invariant, image dependent convolutional propagation and sparsity-aware convolution. This propagation technique is used in a generic cross-domain encoder-decoder architecture with full image guidance at each stage. The network performs sparse-to-dense interpolation for different problems like optical flow, scene flow, depth completion etc., achieving better robustness, accuracy and speed.

FCFR-Net [39] designs a novel end-to-end residual learning framework describing the problem as a two-stage learning task. The coarse-to-fine residual learning framework consists of a sparse-to-dense stage and a coarse-to-fine stage. The former interpolates coarse dense depth map using the CNN framework from [15], while the latter stage further refines the depth maps. A channel shuffle extraction operation is performed to fuse the color and depth features at multi-scale feature levels improving the performance significantly. Also, an energy-based fusion is applied to effectively fuse the features from the channel shuffle stage.

Inspired by FusionNet [24] and DeepLiDAR [20], Hu, Mu, et al. [33] propose a two-branch network PENet, consisting of a color dominant branch and a depth dominant branch. However, the branches are for different purposes and unlike [24] and [20], the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

network can be trained from scratch without requiring any additional datasets. The two branches produce dense depth prediction by exploiting color and depth dominant information. A geometric convolutional layer [52] is used to encode 3D geometric cues. Further, a dilated and accelerated implementation of CSPN++ [23] is proposed to make the refinement more effective and efficient.

Motivated by the popular mechanism of looking and thinking twice in [53], RigNet [34] employs a repetitive design in the image guided network and depth generation branch to gradually and sufficiently recover depth values, resolving the issues related to blurry image guidance and unclear structure in depth. The network consists of a novel repetitive hourglass network, which extracts legible image features of challenging environments to provide more precise guidance for depth recovery. It also uses a repetitive guidance module based on dynamic convolutions [40], including an adaptive fusion mechanism and an efficient guidance algorithm, which can gradually learn precise depth representations.

2.3.2. Guided Image Filtering

Guided Image Filtering is considered another variant of two-branch methods. In the field of depth completion, the idea of guided image filtering refers to the learning and prediction of the kernels from one modality and applying learned kernels to other modalities for feature extraction and fusion.

This approach was first introduced by GuideNet [40]. It proposed a novel method for learning guided kernels from RGB images, applied to depth images to extract features. The intuition is to exploit the properties of guided filtering [54] i.e., spatially variant and content dependent for multi-modal fusion between RGB images and depth maps. However, this is computationally expensive; therefore, it proposes a convolution factorization operation to reduce computation and memory consumption.

Inspired by GuideNet [40], another method has been proposed, which aims to learn steering kernels [55] from RGB images and apply them to sparse depth maps to generate interpolated depth maps [56]. The interpolated depth maps are then refined by utilizing a ResNet [57] to generate the final dense depth maps. The whole pipeline can be trained in an end-to-end manner.

2.3.3. Spatial Propagation Networks (SPN)

The aim of SPN is to learn an affinity matrix to represent the affinities between the pixels. An affinity matrix can be defined as a matrix containing the estimate of the likelihood that pixels (i and j) belong together conditioned on image measurements. The interpretation of the affinity matrix depends on the computer vision task. For instance, in the case of image segmentation task, the affinity matrix should contain semantic-level pairwise similarities.

Depth estimation via affinity learned with convolutional spatial propagation network [58] is one the earliest method , which proposed a generic framework for learning affinity matrix. Instead of manually designing an affinity matrix through similarity kernels for image segmentation, it learned semantic aware affinity values by utilizing deep convolutional neural network (CNN)[59]. Furthermore, the learned affinity matrix is not limited to single computer vision task i.e., image segmentation [60], but it can also be extended to other vision tasks as well. However, it propagates the affinity matrix in a serial fashion, making it inefficient for real-time applications.

Convolutional Spatial Propagation Network (CSPN) [21] extended SPN and presented a convolutional network to learn the affinity matrix for depth completion task. It argues that for a depth refinement task, affinity values of local neighborhood are much more important [21]. To learn the affinity values in local neighborhood, it utilized a deep convolutional neural network and to model long-range context, it uses a recurrent convolutional operation. However, both SPN and CSPN suffers from the problem of fixed local neighborhoods. To counter the problem of fixed local neighborhood in CSPN

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

and SPN, methods including CSPN++ [23], DSPN [61], NLSPN [11] and DySPN [62] were introduced.

CSPN++ [23] added a simple block to CSPN architecture to learn two additional hyper-parameters (1) adaptive convolutional kernel sizes (2) number of iterations for affinity matrix propagation based on image content. Initially, various configurations for both adaptive convolutional kernel sizes and number of iterations for affinity matrix propagation are defined and then during propagation, it learns to predict the correct configuration on each pixel. This leads to significant improvement in both the runtime complexity and the accuracy of depth completion.

Unlike CSPN, DSPN [61] utilized deformable convolutional layers [63] to adaptively generates different receptive field and affinity matrix for each pixel. Later, NLSPN [11] was introduced, which utilized two-stage strategy for depth completion. In the first stage, the proposed method takes RGB and LiDAR sparse depth as an input and outputs (1) non-local neighbors and corresponding affinities of each pixel (2) initial depth estimate (3) confidence map of depth estimate. Then, in second stage, non-local spatial propagation is iteratively performed with confidence-incorporated learnable affinity normalization to generate the final dense depth map. It counters the local affinity problem of CSPN through non-local spatial propagation.

Recently, DySPN [62] propose that instead of using linear propagation for generating affinity matrices, non-linear propagation model should be used for propagation. It dynamically updates the pixel-wise affinity weights by utilizing neighborhood decoupling and spatial-sequential fusion. The neighborhood decoupling is performed by distributing the neighborhood based on the distances between a pixel and its neighborhood and then, recursively generating attention maps based on its propagation stage. Furthermore, it investigates three variants i.e., distance based, dilated [49] and deformable convolutions for determining the optimal number of neighbors required for neighborhood decoupling. Finally, it proposes a diffusion suppression operation to reduce over smoothing of the predicted dense depth maps.

3. Datasets

Typically, depth completion is applied to two kinds of datasets i.e., outdoor and indoor datasets. The outdoor datasets consist of driving sequences, whereas indoor datasets comprise video sequences from a variety of indoor scenes. There exist many such datasets; however, in this paper, we will discuss two famous datasets and benchmarks i.e., KITTI Dataset and its Depth Completion Benchmark (outdoor) [64] and NYU Depth Dataset v2 (indoor) [65], which are used extensively in the field of depth completion for evaluation. The following sections will discuss both KITTI and NYU-v2 datasets in detail.

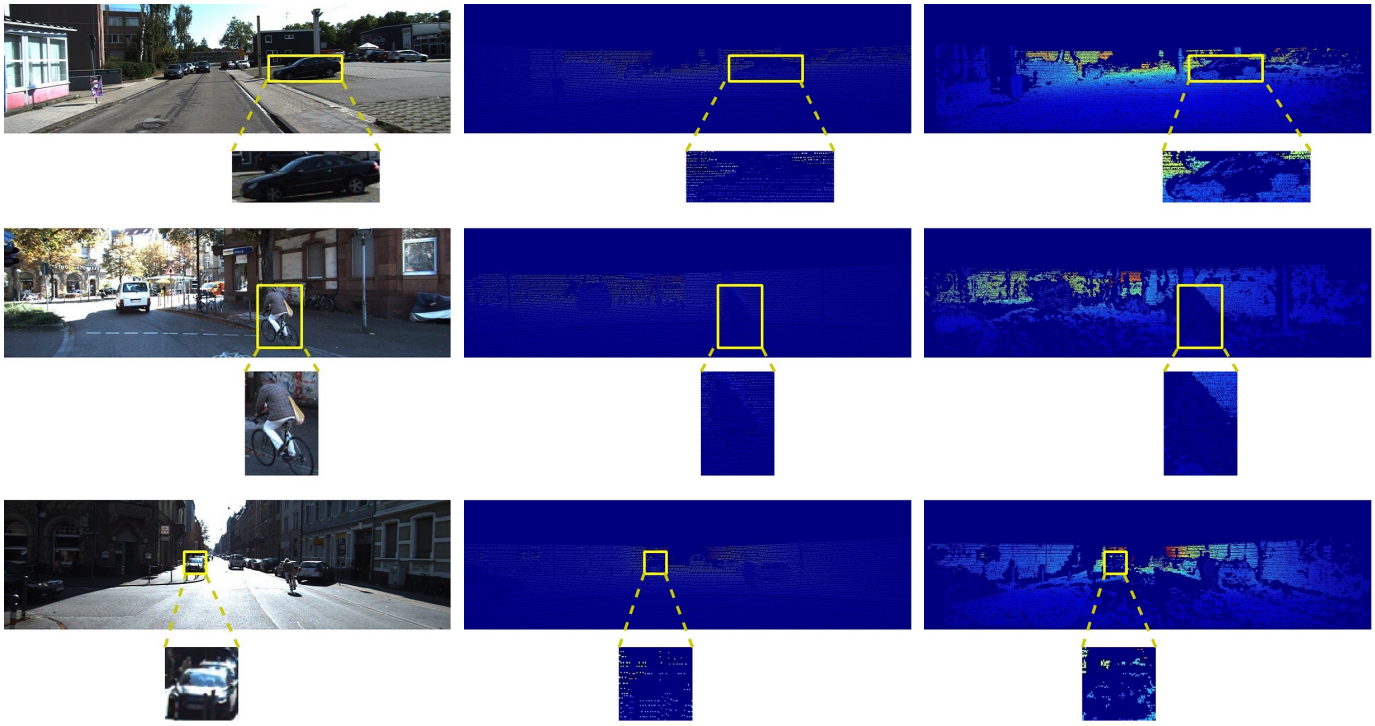
3.1. KITTI Dataset

KITTI dataset [64] is a large outdoor dataset for autonomous vehicles comprising of driving sequences recorded in Karlsruhe, Germany. The driving vehicle VW Passet station is equipped with two stereo camera systems, LiDAR Velodyne HDL-64E laser scanner and an OXTS RT3003 inertial and GPS navigation system. Most of the scenes are collected in rural areas and on the city's highways. The dataset has applications in various computer vision and machine learning research areas, e.g., optical flow, visual odometry, semantic segmentation, semantic instance segmentation, road segmentation, single image depth prediction, depth map completion, 2D and 3D object detection, and object tracking.

3.1.1. KITTI Depth Completion Benchmark

KITTI depth completion [18] benchmark is utilized to evaluate the performance of our approach against existing state-of-the-art methods. It provides 85K sparse depth maps with corresponding RGB images for training, 7K for validation, and 1K for test-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1



(a) RGB Image **(b)** LiDAR sparse Depth Map **(c)** Ground-truth

Figure 7. KITTI depth completion benchmark. Part **a)** shows the aligned RGB images. Part **b)** depicts the sparse LiDAR depth maps, whereas Part **c)** represents the dense ground-truth depth maps. Colorization is applied on LiDAR sparse depth maps and corresponding ground-truth to generate visualizations. The highlighted areas are used to show the sparsity in KITTI depth completion benchmark.

ing. The sparse depth map in the KITTI depth dataset is generated by using LiDAR HDL-64, which provides valid depth values on only 5.9% of all pixels [18,64]. However, the ground-truth contains valid depth values on 16% of all the pixels. The ground-truth is generated by accumulating LiDAR and stereo estimation of the scenes [18,64] using semi-global matching (SGM) [66] approach. Furthermore, the KITTI depth completion dataset also provides an official validation set consisting of 1K frames. Figure 7 presents some images from the depth completion benchmark.

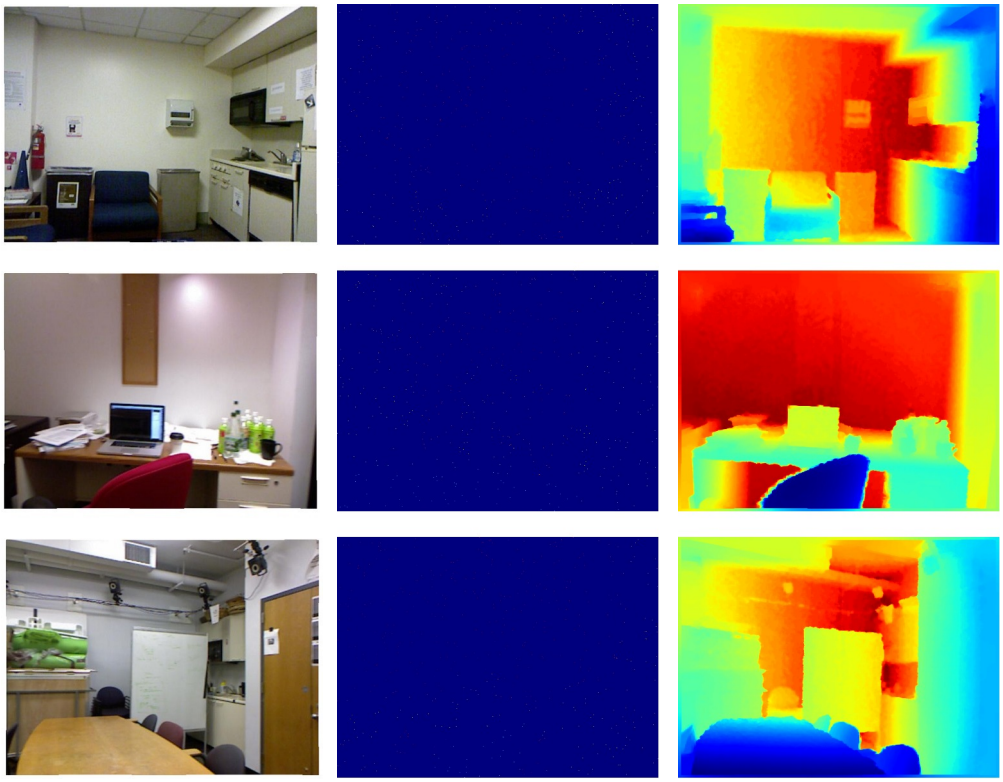
3.2. Nyu-v2 Depth Dataset

It consists of RGB and depth images collected from 464 different indoor scenes. It utilizes a camera to capture RGB data and Microsoft Kinect [67] to record the depth values of the scene. As a preprocessing step, the missing values in depth maps are colorized using a colorized scheme [68]. It provides over 400K images for training; however, most of the methods [33,34,40,44] utilize only a subset for training their approaches. As Kinect provide dense measurements [67], the sparse depth data is generated by randomly removing depth data from the depth ground truth. It also provides 654 images for benchmarking of the results. Figure 8 shows some images from the Nyu-v2 depth dataset.

4. Evaluation Metrics

Depth completion evaluation measures consist of root mean squared error (RMSE), mean absolute error (MAE), root mean squared error of inverse depth (iRMSE), mean absolute error of the inverse depth (iMAE), mean absolute relative error (REL) and threshold accuracy δ. All of the metrics are defined as follows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1



(a) RGB Image **(b)** Kinect sparse Depth Map **(c)** Ground-truth

Figure 8. Nyu-v2 depth dataset. Part **a)** shows the aligned RGB images. Part **b)** depicts the sparse Kinect depth maps, which are generated by randomly sampling only 500 points from the ground truth. Part **c)** represents the fully dense ground-truth depth maps. Colorization is applied on Kinect sparse depth maps and corresponding ground-truth to generate visualizations.

$$RMSE(mm) = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|d_v^{gt} - d_v^{pred}\right|^2} \tag{1}$$

$$MAE(mm) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|d_v^{gt} - d_v^{pred}\right| \tag{2}$$

$$iRMSE(\frac{1}{km}) = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|1/d_v^{gt} - 1/d_v^{pred}\right|^2} \tag{3}$$

$$iMAE(\frac{1}{km}) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|1/d_v^{gt} - 1/d_v^{pred}\right| \tag{4}$$

$$REL(mm) = \frac{1}{\mathcal{V}}\sum_{i=1}^{v}\frac{\left|d_i^{pred} - d_i^{gt}\right|}{d_i^{gt}} \tag{5}$$

$$\delta = \max\left(\frac{d_i^{pred}}{d_i^{gt}}, \frac{d_i^{gt}}{d_i^{pred}}\right) = \delta < \tau, \text{ where } \tau \text{ is the threshold} \tag{6}$$

Among all of the evaluation metrics, RMSE is chosen to rank the submissions on the KITTI and Nyu-v2 Depth online leaderboards.

5. Results

This section compares the results from all the state-of-the art approaches reviewed above. The performance comparison is made both quantitatively and qualitatively. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

quantitative results are reported on the two benchmark datasets for depth completion i.e. KITTI autonomous driving scenes dataset and the NYUv2 indoor scenes dataset. The results on the KITTI dataset are evaluated using the four standard metrics; root mean squared error (RMSE), mean absolute error (MAE), root mean squared error of the inverse depth (iRMSE), and mean absolute error of the inverse depth (iMAE) as shown in Table 1. For the indoor NYUv2 dataset, three metrics are used for evaluation, including the RMSE, mean absolute relative error (REL) and δ_i. Table 2 shows the performance results on the NYUv2 indoor scenes dataset.

Traditional approaches try to directly achieve dense depth maps from sparse depth maps, which causes discontinuities in depth values and loss of structural information. Modern image-guided approaches outperform the traditional ones by a fair margin by using an image as guidance. Spatial propagation methods learn affinity matrices and propagate these to make depth denser. DySPN [62] is the most successful technique in this category and uses non-linear propagation resulting in smoother depth maps. Among the two-branch approaches, RigNet [34] achieves the best results on both the KITTI [64] and NYUv2 [65] datasets. Lastly, GuideNet [40] is the most noticeable work under the guided-kernel depth completion category. Overall, we conclude that two-branch methods show the best results and are currently the state-of-the-art in depth completion. The proper use of multi-modality data allows for the resolution of blurry guidance in images and unclear structure in depth. Also, multi-scale fusion techniques employed by some of the two-branch methods [33,39] prove most successful in extracting discriminate features and fusing them with sparse depth data.

Table 1. Comparison of State-of-the-art approaches on the KITTI Benchmark test dataset. The methods are ordered by their RMSE results from worst to best within each category. The best results within each category are mentioned in bold letters.

Category	Method	RMSE	MAE	iRMSE	iMAE
Two-Branch Networks	SSGP [51]	838.00	245.00	-	-
	DDP [43]	836.00	205.40	2.12	0.86
	MS-Net[LF]-L2 [19]	829.98	233.26	2.60	1.03
	S2D [15]	814.73	249.95	2.81	1.21
	CrossGuidance [48]	807.42	253.98	2.73	1.33
	RSIC [50]	792.80	225.81	2.42	0.99
	Depth-normal [42]	777.05	235.17	2.42	1.13
	FusionNet [24]	772.87	215.02	2.19	0.93
	MSG-CHN [45]	762.19	220.41	2.30	0.98
	DeepLiDAR [20]	758.38	226.50	2.56	1.15
	DenseLiDAR [47]	755.41	214.13	2.25	0.96
	ACMNet [44]	744.91	206.09	2.08	0.90
	FCFR-Net [39]	735.81	217.15	2.20	0.98
	PENet [33]	730.08	210.55	2.17	0.94
	RigNet [34]	**712.66**	203.25	2.08	0.90
Guided Image Filtering	GuideNet [40]	739.24	218.83	2.25	0.99
Spatial Propagation Networks	CSPN [21]	1019.64	279.46	2.93	1.15
	DSPN [61]	766.74	220.36	2.47	1.03
	CSPN++ [23]	743.69	209.28	2.07	0.90
	NLSPN [11]	741.68	199.59	1.99	0.84
	DySPN [62]	**709.12**	192.71	1.88	0.82

6. Conclusion

In this paper, we present a comprehensive survey of depth completion methods. We first present a basic hierarchy of depth completion methodologies consisting of traditional, image-guided, two-branch, spatial propagation networks and guided kernel

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2022 doi:10.20944/preprints202205.0343.v1

Table 2. Comparison of state-of-the-art approaches on the NYUv2 Benchmark dataset. The methods are ordered by their RMSE results from worst to best within each category. The best results within each category are mentioned in bold letters. δ_i denotes the percentage of predicted pixels whose relative error is less than a threshold i (1.25, 1.25^2, and 1.25^3).

Category	Method	RMSE	REL	$\delta_{1.25}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Two-Branch Networks	S2D [15]	0.133	0.027	-	-	-
	EncDec-Net[EF] [19]	0.123	0.017	99.1	99.8	100
	DeepLiDAR [20]	0.115	0.022	99.3	99.9	100.0
	Xu et. al. [42]	0.112	0.018	99.5	99.9	100.0
	FCFR-Net [39]	0.106	0.015	99.5	99.9	100.0
	ACMNet [44]	0.105	0.015	99.4	99.9	100
	DenseLiDAR [47]	0.105	0.015	99.4	99.9	100
	RigNet [34]	**0.090**	0.013	99.6	99.9	100.0
Guided Image Filtering	GuideNet [40]	0.142	0.024	98.8	99.8	100.0
Spatial Propagation Networks	CSPN [21]	0.117	0.016	99.2	99.9	100.0
	CSPN++ [23]	0.116	-	-	-	-
	NLSPN [11]	0.092	0.012	99.6	99.9	100.0
	DySPN [62]	**0.091**	0.012	99.6	99.9	100.0

learning methods. Then, we review the different state-of-the art approaches within each category of the hierarchy by summarizing their contributions and their approach to resolving the prevalent problems of the domain. We further shed light on the most popular benchmark datasets among the research fraternity and the corresponding evaluation metrics reported on each. Finally, to give an overall picture, we present a comparison of all the methods on the discussed benchmarks and reported metrics and concisely mention their pros and cons.

Author Contributions: writing—original draft preparation, M.Z.A.; writing—review and editing, A.U.K, D.N, M.Z.A., H.M.; supervision and project administration, M.L., A.P., D.S. All authors have read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cui, Z.; Heng, L.; Yeo, Y.C.; Geiger, A.; Pollefeys, M.; Sattler, T. Real-time dense mapping for self-driving vehicles using fisheye cameras. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 6087–6093.
2. Häne, C.; Heng, L.; Lee, G.H.; Fraundorfer, F.; Furgale, P.; Sattler, T.; Pollefeys, M. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing* **2017**, *68*, 14–27.
3. Wang, K.; Zhang, Z.; Yan, Z.; Li, X.; Xu, B.; Li, J.; Yang, J. Regularizing Nighttime Weirdness: Efficient Self-supervised Monocular Depth Estimation in the Dark. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16055–16064.
4. Song, X.; Wang, P.; Zhou, D.; Zhu, R.; Guan, C.; Dai, Y.; Su, H.; Li, H.; Yang, R. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5452–5462.
5. Liao, Y.; Huang, L.; Wang, Y.; Kodagoda, S.; Yu, Y.; Liu, Y. Parse geometry from a line: Monocular depth estimation with partial laser observation. 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017, pp. 5059–5066.
6. Dey, A.; Jarvis, G.; Sandor, C.; Reitmayr, G. Tablet versus phone: Depth perception in handheld augmented reality. 2012 IEEE international symposium on mixed and augmented reality (ISMAR). IEEE, 2012, pp. 187–196.
7. Kalia, M.; Navab, N.; Salcudean, T. A Real-Time Interactive Augmented Reality Depth Estimation Technique for Surgical Robotics. 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 8291–8297. doi:10.1109/ICRA.2019.8793610.

8. Holynski, A.; Kopf, J. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (ToG)* **2018**, *37*, 1–11.
9. Armbrüster, C.; Wolter, M.; Kuhlen, T.; Spijkers, W.; Fimm, B. Depth perception in virtual reality: distance estimations in peri-and extrapersonal space. *Cyberpsychology & Behavior* **2008**, *11*, 9–15.
10. Huang, H.C.; Hsieh, C.T.; Yeh, C.H. An Indoor Obstacle Detection System Using Depth Information and Region Growth. *Sensors* **2015**, *15*, 27116–27141. doi:10.3390/s151027116.
11. Park, J.; Joo, K.; Hu, Z.; Liu, C.K.; So Kweon, I. Non-local spatial propagation network for depth completion. European Conference on Computer Vision. Springer, 2020, pp. 120–136.
12. Nguyen, T.N.; Huynh, H.H.; Meunier, J. 3D Reconstruction With Time-of-Flight Depth Camera and Multiple Mirrors. *IEEE Access* **2018**, *6*, 38106–38114. doi:10.1109/ACCESS.2018.2854262.
13. Zhang, Z.; Cui, Z.; Xu, C.; Yan, Y.; Sebe, N.; Yang, J. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4106–4115.
14. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8445–8453.
15. Ma, F.; Cavalheiro, G.V.; Karaman, S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 3288–3295.
16. Chodosh, N.; Wang, C.; Lucey, S. Deep convolutional compressed sensing for lidar depth completion. Asian Conference on Computer Vision. Springer, 2018, pp. 499–513.
17. Jaritz, M.; De Charette, R.; Wirbel, E.; Perrotton, X.; Nashashibi, F. Sparse and dense data with cnns: Depth completion and semantic segmentation. 2018 International Conference on 3D Vision (3DV). IEEE, 2018, pp. 52–60.
18. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity invariant cnns. 2017 international conference on 3D Vision (3DV). IEEE, 2017, pp. 11–20.
19. Eldesokey, A.; Felsberg, M.; Khan, F.S. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *42*, 2423–2436.
20. Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; Pollefeys, M. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3313–3322.
21. Cheng, X.; Wang, P.; Yang, R. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence* **2019**.
22. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
23. Cheng, X.; Wang, P.; Guan, C.; Yang, R. CSPN++: Learning Context and Resource Aware Convolutional Spatial Propagation Networks for Depth Completion. *CoRR* **2019**, *abs/1911.05377*, [1911.05377].
24. Van Gansbeke, W.; Neven, D.; De Brabandere, B.; Van Gool, L. Sparse and noisy lidar completion with rgb guidance and uncertainty. 2019 16th international conference on machine vision applications (MVA). IEEE, 2019, pp. 1–6.
25. Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; Pollefeys, M. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene From Sparse LiDAR Data and Single Color Image. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
26. Bertalmio, M.; Bertozzi, A.L.; Sapiro, G. Navier-stokes, fluid dynamics, and image and video inpainting. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. IEEE, 2001, Vol. 1, pp. I–I.
27. Herrera, D.; Kannala, J.; Heikkilä, J.; others. Depth map inpainting under a second-order smoothness prior. Scandinavian Conference on Image Analysis. Springer, 2013, pp. 555–566.
28. Doria, D.; Radke, R.J. Filling large holes in lidar data by inpainting depth gradients. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012, pp. 65–72.
29. Ferstl, D.; Reinbacher, C.; Ranftl, R.; Rüther, M.; Bischof, H. Image guided depth upsampling using anisotropic total generalized variation. Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 993–1000.
30. Matsuo, K.; Aoki, Y. Depth image enhancement using local tangent plane approximations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3574–3583.
31. Bai, L.; Zhao, Y.; Elhousni, M.; Huang, X. DepthNet: Real-Time LiDAR Point Cloud Depth Completion for Autonomous Vehicles. *IEEE Access* **2020**, *8*, 227825–227833.
32. Eldesokey, A.; Felsberg, M.; Holmquist, K.; Persson, M. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12014–12023.
33. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. Penet: Towards precise and efficient image guided depth completion. 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 13656–13662.
34. Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Xu, B.; Li, J.; Yang, J. RigNet: Repetitive image guided network for depth completion. *arXiv preprint arXiv:2107.13802* **2021**.

35. Zhang, C.; Tang, Y.; Zhao, C.; Sun, Q.; Ye, Z.; Kurths, J. Multitask GANs for Semantic Segmentation and Depth Completion With Cycle Consistency. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*, 5404–5415. doi:10.1109/TNNLS.2021.3072883.
36. Nazir, D.; Liwicki, M.; Stricker, D.; Afzal, M.Z. SemAttNet: Towards Attention-based Semantic Aware Guided Depth Completion, 2022. doi:10.48550/ARXIV.2204.13635.
37. Boulahia, S.Y.; Amamra, A.; Madi, M.R.; Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications* **2021**, *32*, 1–18.
38. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Transactions on Intelligent Transportation Systems* **2022**, *23*, 722–739. doi:10.1109/tits.2020.3023541.
39. Liu, L.; Song, X.; Lyu, X.; Diao, J.; Wang, M.; Liu, Y.; Zhang, L. Fcfr-net: Feature fusion based coarse-to-fine residual learning for monocular depth completion. *arXiv e-prints* **2020**, pp. arXiv–2012.
40. Tang, J.; Tian, F.P.; Feng, W.; Li, J.; Tan, P. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing* **2020**, *30*, 1116–1129.
41. Zhang, Y.; Funkhouser, T. Deep depth completion of a single rgb-d image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 175–185.
42. Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; Li, H. Depth completion from sparse lidar data with depth-normal constraints. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2811–2820.
43. Yang, Y.; Wong, A.; Soatto, S. Dense depth posterior (ddp) from single image and sparse range. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3353–3362.
44. Zhao, S.; Gong, M.; Fu, H.; Tao, D. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing* **2021**, *30*, 5264–5276.
45. Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, C.; others. A multi-scale guided cascade hourglass network for depth completion. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 32–40.
46. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. European conference on computer vision. Springer, 2016, pp. 483–499.
47. Gu, J.; Xiang, Z.; Ye, Y.; Wang, L. DenseLiDAR: A real-time pseudo dense depth guided depth completion network. *IEEE Robotics and Automation Letters* **2021**, *6*, 1808–1815.
48. Lee, S.; Lee, J.; Kim, D.; Kim, J. Deep architecture with cross guidance between single image and sparse lidar data for depth completion. *IEEE Access* **2020**, *8*, 79801–79810.
49. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* **2015**.
50. Yan, L.; Liu, K.; Belyaev, E. Revisiting sparsity invariant convolution: A network for image guided depth completion. *IEEE Access* **2020**, *8*, 126323–126332.
51. Schuster, R.; Wasenmuller, O.; Unger, C.; Stricker, D. Ssgp: Sparse spatial guided propagation for robust and generic interpolation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 197–206.
52. Liu, R.; Lehman, J.; Molino, P.; Such, F.P.; Frank, E.; Sergeev, A.; Yosinski, J. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution, 2018. doi:10.48550/ARXIV.1807.03247.
53. Cao, C.; Liu, X.; Yang, Y.; Yu, Y.; Wang, J.; Wang, Z.; Huang, Y.; Wang, L.; Huang, C.; Xu, W.; others. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. Proceedings of the IEEE international conference on computer vision, 2015, pp. 2956–2964.
54. He, K.; Sun, J.; Tang, X. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *35*, 1397–1409.
55. Tronicke, J.; Böniger, U. Steering kernel regression: An adaptive denoising tool to process GPR data. 2013 7th International Workshop on Advanced Ground Penetrating Radar, 2013, pp. 1–4. doi:10.1109/IWAGPR.2013.6601539.
56. Liu, L.; Liao, Y.; Wang, Y.; Geiger, A.; Liu, Y. Learning steering kernels for guided depth completion. *IEEE Transactions on Image Processing* **2021**, *30*, 2850–2861.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **2015**, *abs/1512.03385*, [1512.03385].
58. Cheng, X.; Wang, P.; Yang, R. Depth estimation via affinity learned with convolutional spatial propagation network. Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 103–119.
59. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems; Pereira, F.; Burges, C.; Bottou, L.; Weinberger, K., Eds. Curran Associates, Inc., 2012, Vol. 25.
60. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* **2015**, *abs/1505.04597*, [1505.04597].
61. Xu, Z.; Yin, H.; Yao, J. Deformable spatial propagation networks for depth completion. 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020, pp. 913–917.
62. Lin, Y.; Cheng, T.; Zhong, Q.; Zhou, W.; Yang, H. Dynamic Spatial Propagation Network for Depth Completion. *arXiv preprint arXiv:2202.09769* **2022**.
63. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773. doi:10.1109/ICCV.2017.89.

64. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 3354–3361.
65. Nathan Silberman, Derek Hoiem, P.K.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. ECCV, 2012.
66. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence* **2007**, *30*, 328–341.
67. Geerse, D.J.; Coolen, B.H.; Roerdink, M. Kinematic Validation of a Multi-Kinect v2 Instrumented 10-Meter Walkway for Quantitative Gait Assessments. *PLOS ONE* **2015**, *10*, 1–15. doi:10.1371/journal.pone.0139913.
68. Levin, A.; Lischinski, D.; Weiss, Y. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*; 2004; pp. 689–694.