Robust Depth Completion with Uncertainty-Driven Loss Functions

Yufan Zhu[1], Weisheng Dong[1]*, Leida Li[1], Jinjian Wu[1], Xin Li[2], Guangming Shi[1]

[1] School of Artificial Intelligence, Xidian University, Xi'an 710071, China
[2] Lane Dep. of CSEE, West Virginia University, Morgantown WV 26506, USA
zhyf70695@163.com, wsdong@mail.xidian.edu.cn, ldli@xidian.edu.cn, jinjian.wu@mail.xidian.edu.cn,
xin.li@mail.wvu.edu, gmshi@xidian.edu.cn

Abstract

Recovering a dense depth image from sparse LiDAR scans is a challenging task. Despite the popularity of color-guided methods for sparse-to-dense depth completion, they treated pixels equally during optimization, ignoring the uneven distribution characteristics in the sparse depth map and the accumulated outliers in the synthesized ground truth. In this work, we introduce uncertainty-driven loss functions to improve the robustness of depth completion and handle the uncertainty in depth completion. Specifically, we propose an explicit uncertainty formulation for robust depth completion with Jeffrey's prior. A parametric uncertain-driven loss is introduced and translated to new loss functions that are robust to noisy or missing data. Meanwhile, we propose a multiscale joint prediction model that can simultaneously predict depth and uncertainty maps. The estimated uncertainty map is also used to perform adaptive prediction on the pixels with high uncertainty, leading to a residual map for refining the completion results. Our method has been tested on KITTI Depth Completion Benchmark and achieved the state-of-the-art robustness performance in terms of MAE, IMAE, and IRMSE metrics.

Introduction

Depth sensing has become increasingly important to a variety of 3D vision tasks, including human-computer interaction (Newcombe et al. 2011), 3D mapping (Zhang and Singh 2014), and autonomous driving (Wang et al. 2019). Depending on the application, 3D sensing of indoor and outdoor environments often faces different technical barriers. For indoor environments, 3D sensing such as Microsoft Kinect and Intel RealSense has become widely affordable, but often suffer from the problem of missing pixels in the presence of shiny/transparent surfaces or inappropriate camera distance (Huang et al. 2019). For outdoor environments, LiDAR is the most popular sensor for acquiring depth information. In addition to the high cost, a fundamental limitation with existing LiDAR sensors is the sparsity of depth measurements. Accordingly, so-called sparse-to-dense completion (a.k.a. depth completion) has been widely studied in the literature (e.g., DeepLiDAR (Qiu et al. 2019), GuideNet (Tang et al. 2020), Adaptive Context-aware Multi-modal Network (Zhao et al. 2020), Non-Local spatial propagation network (Park et al. 2020)). Unlike image super-resolution (SR) (Zhang et al. 2018), the sparse depth map is the degradation (with **non-uniform** downsampling) of distance projection; while the low resolution (LR) RGB image is the **spatially uniform** sampling, which makes the problem of depth completion unique. For instance, we have to handle a significant portion of missing data, as shown in Fig. 1(Top). Traditional full convolution network (FCN) models for SR methods (e.g., RCAN (Zhang et al. 2018), etc.) can not be directly applied to sparse depth maps due to the non-uniform sampling of sparse depth maps as well as the absence of high resolution (HR) RGB images as the ground truth (GT).

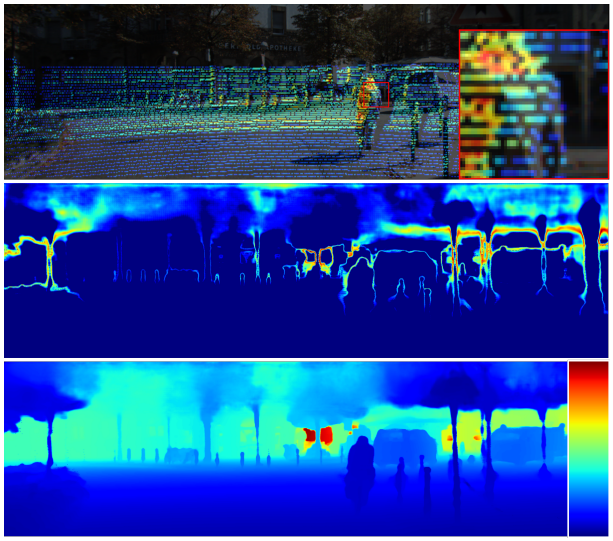

Figure 1: Depth completion with uncertainty prediction. Top: density distribution of a raw Lidar scans with a zoom-in part. Color is proportional to the local sampling density (blue to red is low to high density). To facilitate visual inspection, we have superimposed it on the RGB scene. Middle: the uncertainty map predicted by our method (note that high uncertainty is often associated with depth discontinuities). Bottom: the completed depth image with our method.

Two kinds of uncertainty were introduced in Bayesian deep learning for depth regression and segmentation

*Corresponding author(wsdong@mail.xidian.edu.cn).

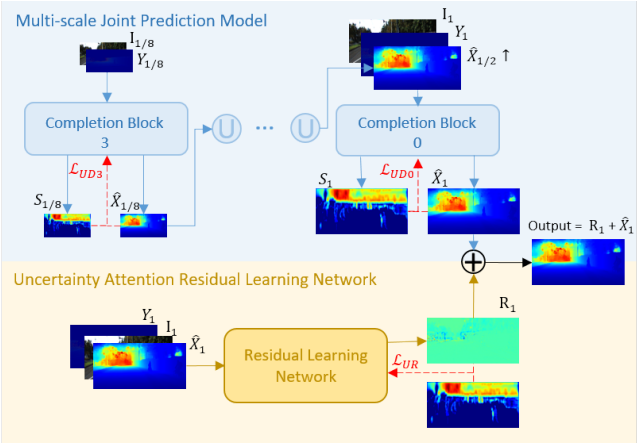

Figure 2: Overview of our method. Top: we jointly predict the uncertainty maps and dense depth images using the Multi-scale Joint Prediction Model. Bottom: Uncertainty Attention Residual Learning Network is used to refine the prediction for pixels of high uncertainty.

(Kendall and Gal 2017). Aleatoric uncertainty captures noise inherent in the observations, and the epistemic uncertainty explains the model uncertainty. In depth completion, aleatoric uncertainty captures noise inherent in raw LiDAR data, which is sparse and unevenly distributed. For example, LiDAR scans the surrounding environment at equally divided angles, resulting in an uneven distribution of depth samples(Uhrig et al. 2017), as shown in Fig. 1. The boundary area of objects and the ultra long-distance areas can hardly be scanned by LiDAR. Such an uneven distribution often leads to the poor prediction of areas with sparse samples. Moreover, the ground truth of KITTI dataset (Geiger et al. 2013) is synthesized by accumulating 11 laser scans and removing outliers by only comparing the depth results in stereo images. Therefore, many outliers will accumulate in semi-sparse GT depth images during synthesis.

In this paper, we introduce the uncertainty-driven loss function to address the uncertainty issue in sparse depth maps and solve the problem of depth completion more effectively. We propose a joint estimation method to simultaneously predict the missing depth values and their uncertainties under a probabilistic framework. By introducing Jeffrey's prior (Figueiredo 2001) to the model, we obtain a new loss function robust to noisy LiDAR data. Inspired by the success of multiscale methods (Shaham, Dekel, and Michaeli 2019; Nah, Hyun Kim, and Mu Lee 2017)), we have carefully designed a multiscale joint prediction network to concurrently estimate the depth and uncertainty maps in a coarse-to-fine manner. The extension of uncertainty-driven loss functions is further complemented by an adaptive prediction of the pixels with high uncertainty, leading to a residual map for refining the depth completion results. The main technical contributions of this work can be summarized as:

- Uncertainty-based deep learning framework for depth completion. For the first time, we propose to develop a more fundamental understanding of aleatoric uncertainty on LiDAR scans for the task of sparse depth completion.
- Uncertainty modeling and estimation. We propose a Multiscale Joint Prediction Model to simultaneously estimate depth and uncertainty maps. The introduction of Jeffrey's prior and multiscale extension both contribute to the improved robustness of our depth completion approach.
- Uncertainty attention residual learning. The aleatoric uncertainty-driven method (Kendall and Gal 2017) was further expanded by residual learning. A new uncertainty-attention network is developed to refine the predicted dense depth with high uncertainty measures.
- Our method has been trained and tested on the popular KITTI benchmark (Geiger et al. 2013). It has achieved the top-ranked performance in terms of MAE, IMAE, and IRMSE metrics among all published papers, which justifies the effectiveness of the proposed approach.

Related Work

Depth estimation from a single image. Depth estimation from a single image (Eigen, Puhrsch, and Fergus 2014) has been extensively studied in the literature. A deep structured learning scheme was proposed in (Liu, Shen, and Lin 2015; Long, Shelhamer, and Darrell 2015) to learn the unary and pairwise potentials of continuous conditional random field (CRF) by a unified deep convolutional neural network (CNN) framework. This line of research was extended into single-image depth estimation in the wild (Chen et al. 2016) - i.e., recovering the depth from a single image taken in unconstrained settings. More recently, a spacing-increasing discretization (SID) strategy was introduced in (Fu et al. 2018) to discretize depth values and recast depth estimation as an ordinal regression problem.

Model-based sparse-to-dense depth completion methods. Early methods of estimating dense measurements from sparse ones have been considered under the framework of compressed sensing. For example, (Hawe, Kleinsteuber, and Diepold 2011) shows disparity maps can be reconstructed from only a small set of reliable support points based on the compressive sensing principle. In (Liu, Chan, and Nguyen 2015), a combined dictionary of wavelets and contourlets proved to improve the reconstruction quality of disparity maps. In (Ku, Harakeh, and Waslander 2018), a simple and fast method was developed to complete a sparse depth map using basic image processing operations only.

Deep learning based depth completion. By designing the U-net and fusing the aligned color image, self-supervised approaches such as (Ma, Cavalheiro, and Karaman 2019) have achieved significant improvement over traditional model-based methods. In (Van Gansbeke et al. 2019), a new framework was proposed to extract both global and local information from two different networks during depth completion. More recently, (Tang et al. 2019) introduced the guided convolution module into a network architecture to fuse color information and developed spatially separable convolution to reduce computational complexity. There are also methods(Qiu et al. 2019; Chen et al. 2019) using auxiliary variables to assist the task of depth completion. In (Qiu et al. 2019), surface normals are estimated from color images and

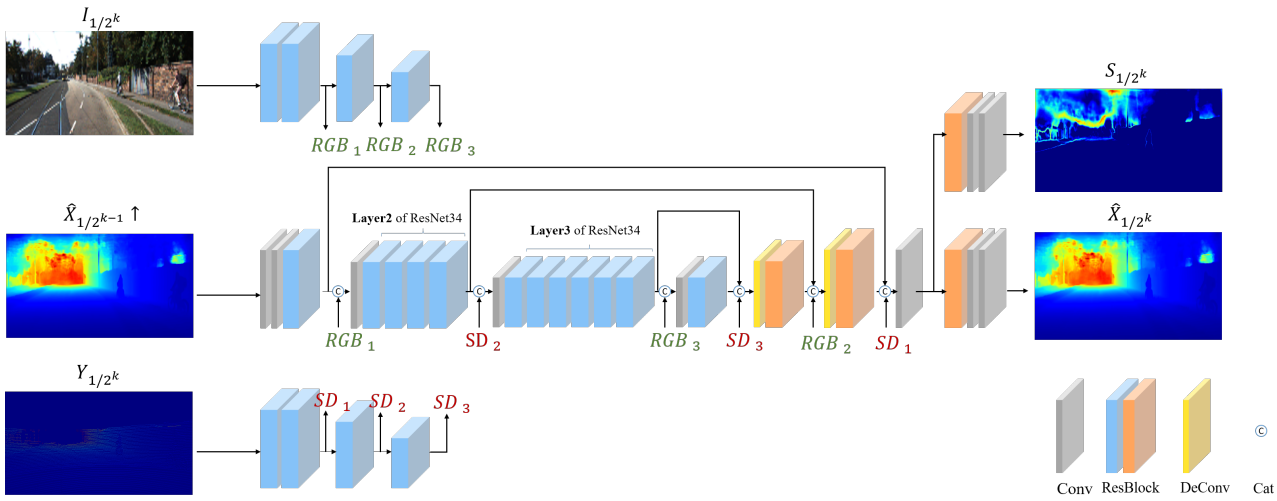

Figure 3: A Depth Completion Block in Multi-scale Joint Prediction Network. Input the up-sampling results of the previous module to generate fine-dense depth images by fusing with RGB and sparse depth.

imposed as a geometric constraint for depth completion. In (Chen et al. 2019), sparse depth maps are converted to 3D point clouds, then both 2D and 3D features are exploited to improve the performance of depth completion.

Uncertainty estimation. Early work(MacKay 1992; Neal 2012) like Bayesian neural networks (BNNs) has been designed to obtain uncertainty estimates. In the work(Kendall and Gal 2017), the aleatoric uncertainty from the noise in the observations and the epistemic uncertainty explaining the model uncertainty were studied in a joint framework in the tasks like pixel-wise depth regression and semantic segmentation. And Recently, more methods (Qiu et al. 2019; Van Gansbeke et al. 2019; Xu et al. 2019) promoted the prediction of Uncertainty to filter out noisy predictions within the network.(Van Gansbeke et al. 2019; Xu et al. 2019) predicts the confidence(just like uncertainty) to fuse the completed depth differently, like local and global networks. (Qiu et al. 2019) learned confidence mask to represent the noisy depth of occluded regions. More recently, researchers in (Chang et al. 2020) have investigated the data uncertainty with estimated mean and variance in face recognition. (Wong and Soatto 2019; Wong et al. 2021) developed adaptive regularization and data fidelity to handle hard boundary prediction in unsupervised depth completion.

Approach

We first propose a probabilistic method for joint training of depth and uncertainty maps. Then we present how to implement our proposed loss functions in a multiscale network to predict the depth and uncertainty maps from coarse to fine, which is named the Uncertainty-based Multi-scale Joint Prediction Model. Finally, we show how to use the predicted uncertainty map to improve the result of depth completion by an Uncertainty Attention Residual Learning Network.

Uncertainty-Driven Loss Function

In low-level vision tasks, Bayesian deep learning offers a principled framework for taking uncertainty into account (Kendall and Gal 2017). In Bayesian modeling, there are primarily two types of uncertainty: *aleatoric* and *epistemic*. Aleatoric uncertainty capturing noise inherent in the observations can be further categorized into two classes: homoscedastic and heteroscedastic. Under the context of depth completion, heteroscedastic uncertainty is especially important due to the physical limitations of LiDAR sensors. For example, LiDAR scans the surrounding environment at equally divided angles, resulting in an uneven distribution of depth images (Uhrig et al. 2017); such an uneven distribution often leads to varying difficulties in different densities areas.

In previous depth completion methods (Ma, Cavalheiro, and Karaman 2019; Qiu et al. 2019; Chen et al. 2019), the MSE Loss was used to average the errors across all pixels. Low-density areas (arising from non-uniform sampling) and outliers often cause the network to over-emphasize these areas (i.e., overfitting). Inspired by the recent work of uncertainty modeling (Lee and Chung 2019), we consider a parametric approach of quantifying the uncertainty in a depth map by its variance field Σ. First, we define the sparse depth image as Y, the corresponding dense depth image (GT) as X, and the process of generating dense depth images based on the deep learning network is $X = F(Y)$, where we expect the predicted $\hat{X}$ to approximate X. Thus, the procession of depth completion can be expressed as maximizing the posterior probability $P(X|Y)$. By introducing the uncertainty measure Σ (σ for a pixel), we can decompose the joint posterior probability into the product of marginals:

$$\begin{aligned} P(X, \Sigma|Y) =& P(\Sigma|Y)\, P(X|\Sigma, Y) \\ =& \prod p(\sigma_i|y_i)\, p(x_i|\sigma_i, y_i) \end{aligned} \tag{1}$$

where x_i, σ_i, y_i is the pixel-wise element of X, Σ, and Y.

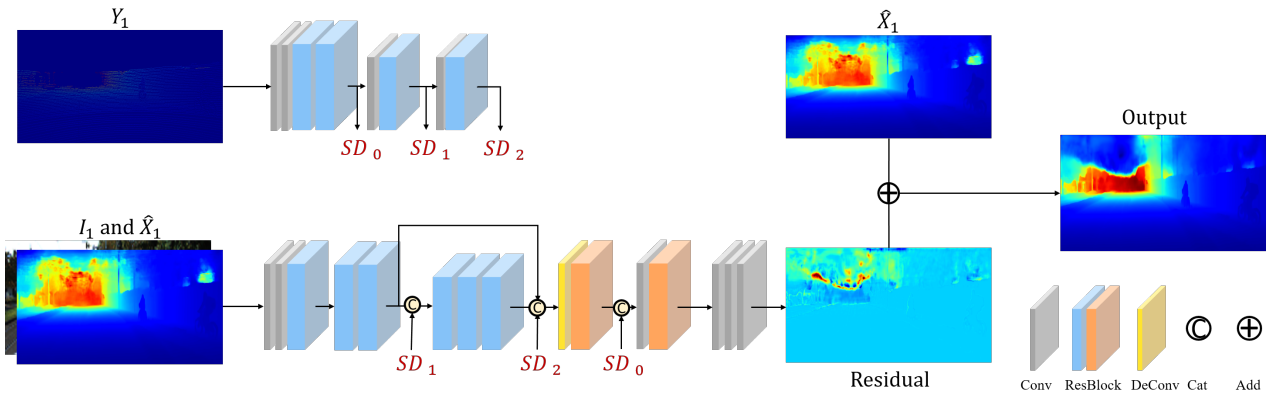

Figure 4: Uncertainty Attention Residual Learning Network.

For the likelihood of uncertainty map $p(\sigma_i|y_i)$, we model it with the Jeffrey's prior $P(\sigma_i|y_i) \approx \frac{1}{\sigma_i}$ (Figueiredo 2001) based on the intuition of the sparsity on uncertainty map. For the likelihood term, $p(x_i|\sigma_i, y_i)$ can be modeled by a Gaussian distribution observing $N(x_i, \sigma_i)$:

$$p(x_i|\sigma_i, y_i) \approx \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(\hat{x}_i - x_i)^2}{2\sigma_i^2}) \tag{2}$$

where $\hat{x}_i$ is a pixel of $\hat{X}$. Therefore, we obtain the following MAP problem:

$$\begin{aligned} &\max \sum (\log p(\sigma_i|y_i) + \log p(x_i|\sigma_i, y_i)) \\ &= \underset{\hat{x}_i, \sigma_i}{\operatorname{argmax}} \sum \left(-2\log\sigma_i - \frac{(\hat{x}_i - x_i)^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi \right) \\ &= \underset{\hat{x}_i, \sigma_i}{\operatorname{argmin}} \sum \left(4\log\sigma_i + \frac{(\hat{x}_i - x_i)^2}{\sigma_i^2} \right) \\ &= \underset{\hat{x}_i, s_i}{\operatorname{argmin}} \sum \left(e^{-s_i} (\hat{x}_i - x_i)^2 + 2s_i \right) \end{aligned} \tag{3}$$

where $s_i = 2\log\sigma_i$, $\sigma_i^2 = e^{s_i}$. Such formulation of uncertainty modeling can be translated to the design of a new loss function. Conceptually similar to the previous work (Kendall and Gal 2017), we can build the final optimized loss, named Uncertainty-Depth Joint Optimization Loss Function as follows:

$$\mathcal{L}_{UD} = \frac{1}{N} \sum \left(e^{-s_i} (\hat{x}_i - x_i)^2 + 2s_i \right) \tag{4}$$

From the formula, we conclude that the first term will reduce the joint loss of pixels with large differences between the prediction and ground truth $(\hat{x}_i - x_i)^2$. During the process of optimization, the optimizer may increase the uncertainty values so large that the penalty term e^{-s_i} eventually approaches zero. To balance the first term, the second term limits the growth of uncertainty s_i as a regularization term. As a consequence of balancing, the network will control the contribution of high-uncertainty regions to the joint loss function, instead of overfitting these regions. In summary, Eq. (4) makes the overall network focused more on easy-to-predict regions by attenuating hard-to-predict regions to strike an improved tradeoff.

Multiscale Joint Prediction Model

We present how to implement the joint uncertainty-depth optimization model. Our network implementation contains two basic modules: the Multi-scale joint prediction network as shown in Fig. 2(Top), the Sparse-to-dense Basic Completion Block as shown in Fig. 3.

Multiscale Network Structure. Similar multiscale architecture has been shown effective for various low-level vision tasks such as image synthesis (Shaham, Dekel, and Michaeli 2019) and image deblurring (Nah, Hyun Kim, and Mu Lee 2017). The multiscale module starts with a pair of downsampled color image $I_{1/2^k}$(Interpolation) and depth map $Y_{1/2^k}$(Max-Pooling), where k is the downsampling factor. To handle sparsity, we use a relatively large-size convolution kernel to extract and propagate spatial features from the coarse to fine scale. Throughout a series of upsampling operations after completion, the size of the convolution kernel is kept constant, but the size of image patches (receptive field) decreases relatively to rich more details. Unlike previous works, we have found that an interpolation upsampling layer in the pixel domain between completion blocks, as shown in Fig. 2, is preferred for coarse-to-fine progression over an upconvolution layer, which has strong local constraints than the deconvolution process (often used for superresolution in the feature domain (Sajjadi, Scholkopf, and Hirsch 2017)). Meanwhile, such a multiscale framework allows us to selectively supervise the output(depth and uncertainty) of varying patch sizes and adaptively calculate the loss function for performance optimization. To optimize the uncertainty and depth values at all scales at the same time, we define the following weighted multi-objective optimization function.

$$\mathcal{L}_{StepOne} = \sum \omega_k \mathcal{L}_{UDk} \tag{5}$$

where ω_k and $\mathcal{L}_{UDk}$ denote the weighting coefficient and loss term at the k-th scale, respectively.

Sparse-to-dense Basic Completion Block. The completion block's backbone is constructed based on the well-known U-net that has shown effectiveness in image segmentation tasks (Ronneberger, Fischer, and Brox 2015). In our construction, the backbone of the model consists of two branches, each of which will output the depth image and uncertainty map

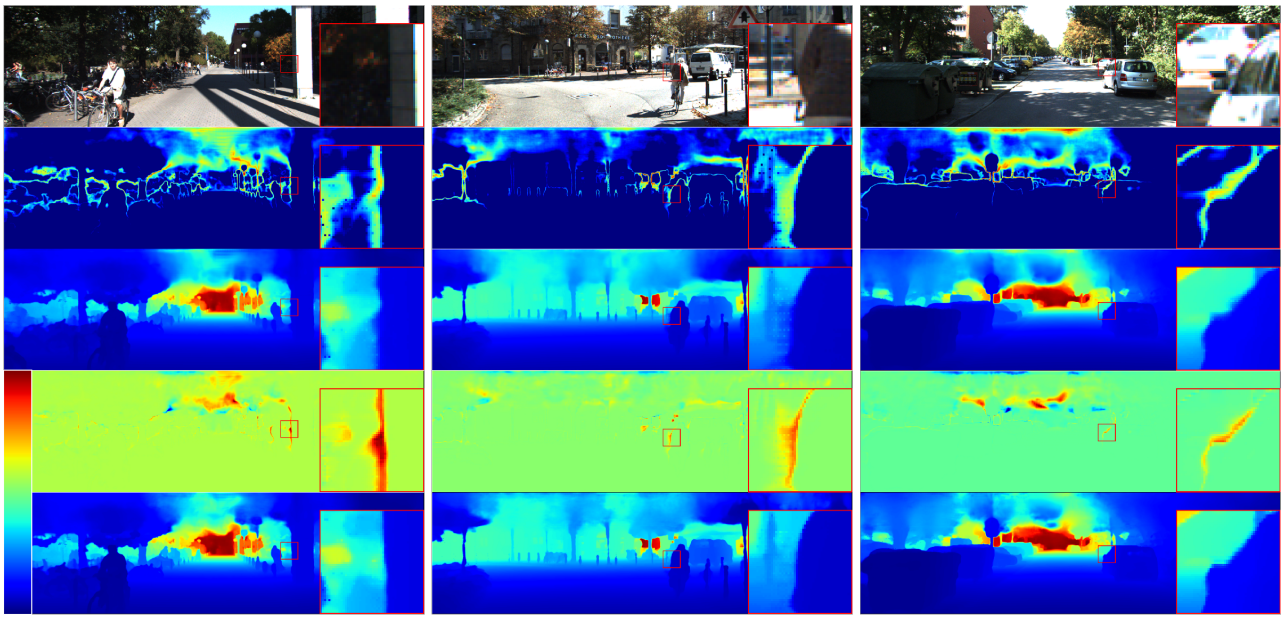

Figure 5: Demonstration of depth completion results after each stage. First row: the color images. Second row: the uncertainty maps. Third row: the dense depth images generated by Multi-scale Joint Prediction Model. Fourth row: the residual maps generated by Uncertainty Attention Residual Learning Network. Fifth row: the final output images.

respectively. Similar architecture has also been used in self-supervised depth completion (Ma, Cavalheiro, and Karaman 2019) where it takes the layers of original Resnet34 (He et al. 2016) as the encoder and some deconvolution layers of (Zeiler, Taylor, and Fergus 2011) as the decoder. By contrast, our design reflects the unique strategy of fusing color and depth information in an alternating manner, as marked by the bars with different colors and heights in Fig. 3. Additionally, we have added some skip connections to the residual modules to further facilitate the information flow between deconvolution operations (Van Gansbeke et al. 2019).

Uncertainty Attention Residual Learning Network

Through the previous analysis, we know that the model has achieved an improvement in the overall recovery performance by alleviating those difficult-to-recover areas (regions with great uncertainty). From the predicted uncertainty map in Fig. 5, we can observe that the edge regions of the object and the regions with larger depth values often have larger uncertain values. Note that the optimization constraints of these regions are relatively relaxed in the first step (Multi-scale Joint Prediction). The new insight behind our residual learning network in the second step is to use the estimated uncertainty map to guide the procedure of depth completion *refinement*. In other words, with the knowledge about the distribution of high-uncertainty regions, we hope to tailor the process of optimization for these special regions to achieve an even better completion result.

Our key idea is to predict the refinement map R for $\hat{X}_1$, only for the pixels that are uncertain in the first step. Note that the predicted uncertainty map S_1 is used to give higher weight to the loss in the regions of high uncertainty. Based on the above reasoning, the loss function associated with residue learning can be formed by:

$$\begin{aligned} \mathcal{L}_{UR} &= \frac{1}{N}\sum s_i \left| \left(x_i - \hat{x}_i\right) - r_i \right| \\ \mathcal{L}^2_{UR} &= \frac{1}{N}\sum s_i \left(\left(x_i - \hat{x}_i\right) - r_i \right)^2 \end{aligned} \tag{6}$$

where r_i is the pixel of predicted residual R, $\hat{x}_i$ is the depth output of Multiscale Joint Prediction Network and we will use a mixture of L1 and L2 forms to build the epoch-dependent balanced loss function next. The structure of the network is a simplified U-net, which takes RGB images and $\hat{X}_1$ as input, incorporates the information of the sparse depth in the prediction process, and finally outputs the residual image. The final output of depth completion is $R + \hat{X}_1$, as shown in Fig. 4.

It is worth mentioning that the optimization of different objective metrics for depth completion often has conflicting objectives. For example, the objective of minimizing RMSE is often inconsistent with that of minimizing MAE (the former is more sensitive to outliers than the latter). Note that RMSE is equivalent to the square root of L2 loss and MAE is equivalent to the L1 loss; they respectively characterize different performance metrics for the task of depth completion. A compromised solution is to formulate a balanced loss function between different metrics. In our current implementation, we have used the following epoch-dependent balanced loss function for the residual learning network:

$$\mathcal{L}_{StepTwo} = \mathcal{L}_{URB} = \begin{cases} \mathcal{L}_{UR}, & N_{epoch}\ is\ even \\ \frac{1}{2}\left(\mathcal{L}_{UR} + \mathcal{L}^2_{UR}\right), & else \end{cases} \tag{7}$$

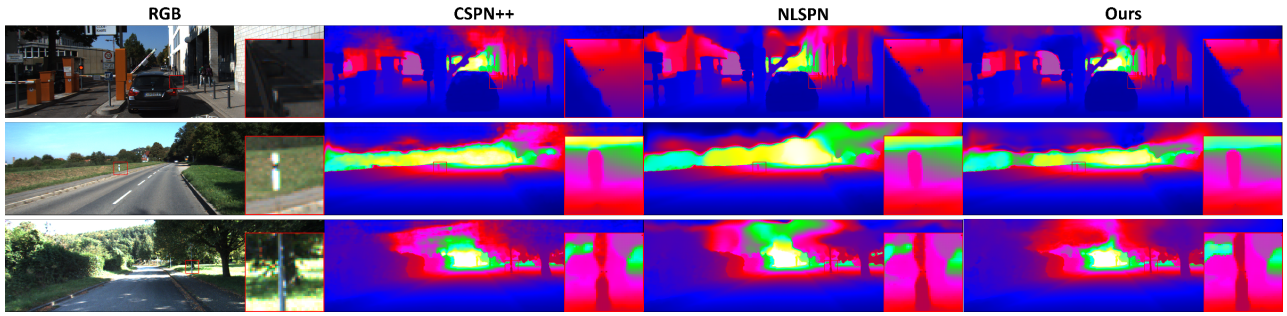

Figure 6: Visual quality comparison on KITTI Test Benchmark(Uhrig et al. 2017). Left to right: RGB, CSPN++(Cheng et al. 2020), NLSPN(Park et al. 2020), Our results.

Methods	MAE	IMAE	RMSE	IRMSE
NLSPN	199.59	0.84	741.68	1.99
GuideNet	218.83	0.99	736.24	2.55
CSPN++	209.28	0.90	743.69	2.07
DeepLiDAR	226.50	1.15	758.38	2.56
Sparse-to-Dense (gd)	249.95	1.21	814.73	2.80
RGB_guide&certainty	215.02	0.93	772.87	2.19
Ours(with L_{URB})	198.09	0.85	751.59	**1.98**
Ours(with L_{UR})	**190.88**	**0.83**	795.43	**1.98**

Table 1: Comparison with other SOTA methods on KITTI Test Benchmark.

In summary, this epoch-dependent loss function aims at better balancing the RMSE and MAE metrics by adaptively combining L1 and L2 losses. The benefit of using a hybrid yet balanced loss function is shown in Table 5.

Experiments

Experimental Settings

Dataset. The KITTI depth completion benchmark(Uhrig et al. 2017) has 86898 Lidar frames for training, 1000 frames for validation, and 1000 frames for testing. Each frame has one depth map from the LiDAR sensor and one aligned color image from the visible spectrum camera.

Evaluation Metrics. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Inverse RMSE (IRMSE), and inverse MAE (IMAE).

Parameter Settings. Our training is implemented by Pytorch with 5 NVIDIA GTX2080Ti GPUs and set batch-size to 5. In our current implementation, we have used ADAM (Kingma and Ba 2014) as the optimization algorithm. We have set the learning rate to 1×10^{-4} when we train our multiscale joint prediction model and 2×10^{-4} when training uncertainty attention residual learning model. The acceleration of learning rate will decline as the epoch increases. The other parameters are all the same with $(\beta_1,\ \beta_2)$ = $(0.9, 0.999)$, $eps = 1 \times 10^{-8}$ and $Weight\_decay = 0$.

Mask Loss. Since the groundtruth is semi-dense, a binary mask is applied to the loss function only accounting for the valid depth points.

Unc	Jef	Res	MAE	IMAE	RMSE	IRMSE
×	×	×	211.01	0.92	792.34	2.185
✓	×	×	203.94	0.86	791.13	2.048
✓	✓	×	204.61	0.87	783.87	2.072
✓	✓	✓	186.90	0.81	833.67	2.069

Table 2: Ablation study of three key modules on Val set (uncertainty-driven loss, Jeffrey's prior, and residue learning).

Performance Comparison

Comparison with SOTA methods. We have compared our method with the SOTA methods on the KITTI TEST benchmark. Our method has achieved highly competitive results in terms of all metrics. As shown in Table 1, it surpasses all other competing methods on MAE, IMAE, and IRMSE metrics, which demonstrates the superiority of our method. We have also shown some qualitative visual comparison results on the test dataset of KITTI depth completion benchmark in Fig. 6. Our results have clear boundaries and recover more details than other methods. For example, in the first row, our method is the only one capable of restoring the rearview mirror hidden in the dark background; in the third row, the vertical pole is better recovered by our method.

Ablation Studies and Analyses

We have tested the influence of several main modules on the experimental results. As shown in Table 2, all three modules have jointly contributed to the performance improvement. However, the impact of different module varies on different quality metrics - e.g., MAE and IMAE are more consistent with each other, while RMSE is not. It can be verified that introduction of the Jeffrey prior greatly boosts the RMSE performance, which validates its effectiveness.

We have also experimented on Val set with the effect of different numbers of depth completion blocks (**NS** in the Table 3), and found NS = 4 reached the best metrics (3 out of 4) when the smallest downsampled image size is $(152, 44)$(hard to be downsampled for U-net). In addition to using the sparse depth images as input correction information in Uncertainty Attention Residual Learning Network

NS	MAE	IMAE	RMSE	IRMSE
1	215.06	0.93	786.02	2.187
2	205.95	0.88	778.05	2.110
3	206.46	0.88	782.18	2.113
4	204.61	0.87	783.87	2.072

Table 3: Ablation experiment on the Number (NS) of S-to-D Basic Completion Block in Multiscale Joint Prediction.

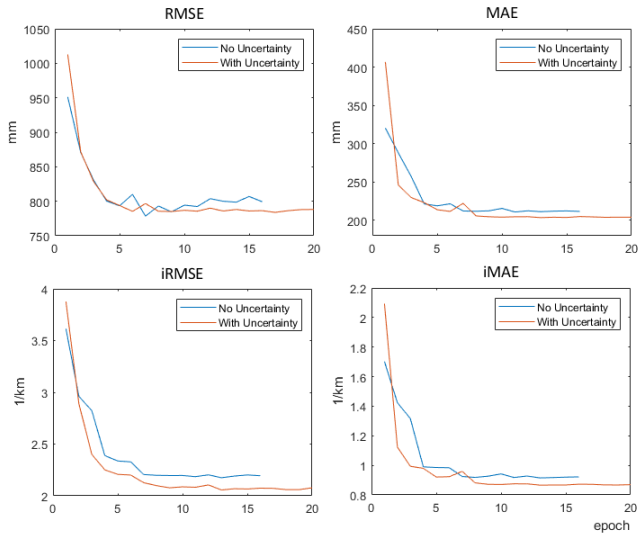

Figure 7: Convergence analysis. Training with uncertainty loss converges faster and reach lower bound than w/o.

(Table 4), we have tested the effects of different input combinations on the results. It turns out that the combination of $\hat{X}_1$ and RGB image has achieved the best performance.

Convergence and Qualitative Analysis

Multiscale Joint Prediction Model. This module will jointly predict the uncertainty map with dense depth image. From Table 2 and Fig. 7, we can observe that better results and robustness can be achieved by uncertainty-driven loss. By visually inspecting the uncertainty map in Fig. 5, we clearly see that the areas with high uncertainty are concentrated in the area of the object boundary (the depth drop is large) and the area with the higher depth value (distant road or open area). In Fig. 7. Our empirical studies have shown that uncertainty-driven loss tends to produce a more chaotic start, but the optimization of the network will converge faster, and the objective metrics will stabilize around the optimal values after the convergence.

$\hat{X}_1$	I_1(RGB)	MAE	IMAE	RMSE	IRMSE
×	✓	196.48	0.847	791.01	2.066
✓	×	187.31	0.809	834.33	2.037
✓	✓	186.90	0.807	833.67	2.069

Table 4: Ablation experiment on the input of Uncertainty Attention Residual Learning module.

Residual	Loss	MAE	IMAE	RMSE	IRMSE
×	L_{UD}	204.61	0.87	783.87	2.072
✓	L_{UR}	186.90	0.81	833.67	2.069
✓	L_{URB}	195.09	0.83	786.73	2.018

Table 5: Ablation experiment on the loss function of Uncertainty Attention Residual Learning.

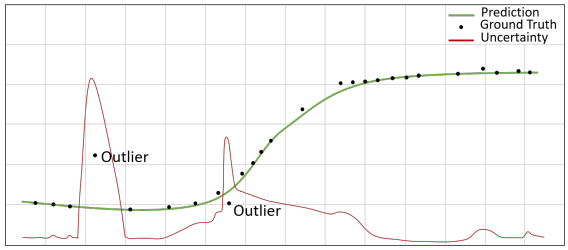

Figure 8: A qualitative example for the uncertainty in our method. Areas with outliers, sparse, or large changes in depth values have higher uncertainty values, and these areas are weighted less during the joint training process.

Uncertainty Attention Residual Learning Network. We reuse the uncertainty map to focus on the part that contains more uncertainty in the multiscale joint prediction model. Table 2 and 5 verify how the prediction of the residual map boosts the performance. Visual inspection of the residual map in Fig. 5 shows that most areas are close to zero (light blue) and high uncertainty areas are corrected (red is increasing and blue is decreasing). For a more intuitive explanation, please see a concrete example as shown in Fig. 8. It clearly demonstrates that outliers are assigned high uncertainty values and therefore get attenuated by the joint training process.

Conclusion

We introduce uncertainty into depth completion and propose how to improve the estimation for areas with high uncertainty. We propose a probabilistic Joint Training Method with Jeffrey's prior, consisting of the Multi-scale Joint Prediction Network and the Uncertainty Attention Residual Learning Network. Our method overcomes the difficulties caused by the uneven distribution and outliers in both LiDAR scanned depth images and synthesized semi-dense images. Extensive experimental results are reported to show that our method has better performance than existing top-rank published methods on the KITTI depth completion benchmark. Meantime, the computational complexity of our method is the lowest among the top methods - the actual running time of only 0.07s per frame can meet the requirement of real-time processing in practical applications. In addition, the related structure of uncertainty prediction can be removed totally after training, which will reduce network parameters further. In the future, we plan to study multi-modal data fusion with LiDAR and other sensors.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0101400 and the Natural Science Foundation of China under Grant 61991451, Grant 61632019, Grant 61621005, and Grant 61836008. Xin Li's work is partially supported by the NSF under grants IIS-1951504 and OAC-1940855, the DoJ/NIJ under grant NIJ 2018-75-CX-0032, and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5).

References

Chang, J.; Lan, Z.; Cheng, C.; and Wei, Y. 2020. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5710–5719.

Chen, W.; Fu, Z.; Yang, D.; and Deng, J. 2016. Single-image depth perception in the wild. In *Advances in neural information processing systems*, 730–738.

Chen, Y.; Yang, B.; Liang, M.; and Urtasun, R. 2019. Learning joint 2d-3d representations for depth completion. In *Proceedings of the IEEE International Conference on Computer Vision*, 10023–10032.

Cheng, X.; Wang, P.; Guan, C.; and Yang, R. 2020. CSPN++: Learning Context and Resource Aware Convolutional Spatial Propagation Networks for Depth Completion. In *AAAI*, 10615–10622.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, 2366–2374.

Figueiredo, M. A. 2001. Adaptive Sparseness Using Jeffreys Prior. In *NIPS*, 697–704.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2002–2011.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Hawe, S.; Kleinsteuber, M.; and Diepold, K. 2011. Dense disparity maps from sparse disparity measurements. In *2011 International Conference on Computer Vision*, 2126–2133. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.-K.; Wu, T.-H.; Liu, Y.-C.; and Hsu, W. H. 2019. Indoor Depth Completion with Boundary Consistency and Self-Attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ku, J.; Harakeh, A.; and Waslander, S. L. 2018. In defense of classical image processing: Fast depth completion on the cpu. In *2018 15th Conference on Computer and Robot Vision (CRV)*, 16–22. IEEE.

Lee, C.; and Chung, K.-S. 2019. GRAM: Gradient rescaling attention model for data uncertainty estimation in single image super resolution. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 8–13.

Liu, F.; Shen, C.; and Lin, G. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5162–5170.

Liu, L.; Chan, S. H.; and Nguyen, T. Q. 2015. Depth Reconstruction From Sparse Samples: Representation, Algorithm, and Sampling. *IEEE Transactions on Image Processing*, 24(6): 1983–1996.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Ma, F.; Cavalheiro, G. V.; and Karaman, S. 2019. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *2019 International Conference on Robotics and Automation (ICRA)*, 3288–3295. IEEE.

MacKay, D. J. 1992. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3): 448–472.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Neal, R. M. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Newcombe, R. A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A. J.; Kohi, P.; Shotton, J.; Hodges, S.; and Fitzgibbon, A. 2011. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, 127–136. IEEE.

Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; and Kweon, I. S. 2020. Non-Local Spatial Propagation Network for Depth Completion. *arXiv preprint arXiv:2007.10042*.

Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; and Pollefeys, M. 2019. 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*, 3313–3322.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sajjadi, M. S.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 4491–4500.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4570–4580.

Tang, J.; Tian, F.-P.; Feng, W.; Li, J.; and Tan, P. 2019. Learning guided convolutional network for depth completion. *arXiv preprint arXiv:1908.01238*.

Tang, J.; Tian, F.-P.; Feng, W.; Li, J.; and Tan, P. 2020. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30: 1116–1129.

Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; and Geiger, A. 2017. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, 11–20. IEEE.

Van Gansbeke, W.; Neven, D.; De Brabandere, B.; and Van Gool, L. 2019. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th International Conference on Machine Vision Applications (MVA)*, 1–6. IEEE.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8445–8453.

Wong, A.; Fei, X.; Hong, B.-W.; and Soatto, S. 2021. An Adaptive Framework for Learning Unsupervised Depth Completion. *IEEE Robotics and Automation Letters*, 6(2): 3120–3127.

Wong, A.; and Soatto, S. 2019. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5644–5653.

Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; and Li, H. 2019. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, 2811–2820.

Zeiler, M. D.; Taylor, G. W.; and Fergus, R. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, 2018–2025. IEEE.

Zhang, J.; and Singh, S. 2014. LOAM: Lidar Odometry and Mapping in Real-time. *Robotics: Science and Systems*, 2(9).

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.

Zhao, S.; Gong, M.; Fu, H.; and Tao, D. 2020. Adaptive context-aware multi-modal network for depth completion. *arXiv preprint arXiv:2008.10833*.