Learned Semantic Multi-Sensor Depth Map Fusion

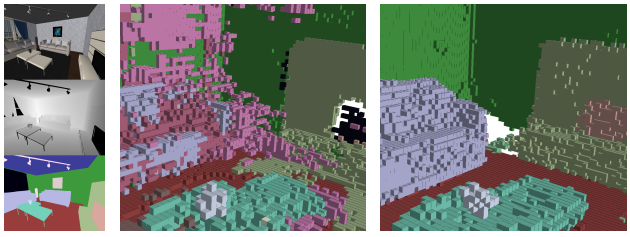
Denys Rozumnyi[1,3] Ian Cherabier[1] Marc Pollefeys[1,2] Martin R. Oswald[1]
[1]Department of Computer Science, ETH Zurich [2]Microsoft
[3]Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

Abstract

Volumetric depth map fusion based on truncated signed distance functions has become a standard method and is used in many 3D reconstruction pipelines. In this paper, we are generalizing this classic method in multiple ways: 1) Semantics: Semantic information enriches the scene representation and is incorporated into the fusion process. 2) Multi-Sensor: Depth information can originate from different sensors or algorithms with very different noise and outlier statistics which are considered during data fusion. 3) Scene denoising and completion: Sensors can fail to recover depth for certain materials and light conditions, or data is missing due to occlusions. Our method denoises the geometry, closes holes and computes a watertight surface for every semantic class. 4) Learning: We propose a neural network reconstruction method that unifies all these properties within a single powerful framework. Our method learns sensor or algorithm properties jointly with semantic depth fusion and scene completion and can also be used as an expert system, e.g. to unify the strengths of various photometric stereo algorithms. Our approach is the first to unify all these properties. Experimental evaluations on both synthetic and real data sets demonstrate clear improvements.

1. Introduction

Holistic 3D scene understanding is one of the central goals of computer vision research. Tremendous progress has been made within the last decades to recover accurate 3D scene geometry with a variety of sensors [8, 23, 35] and image-based 3D reconstruction methods [19, 49, 39]. With the breakthrough in machine learning, algorithms that recover 3D geometry increasingly include semantic information [26, 20, 21, 1, 5, 31, 10, 14, 12, 6, 43] in order to improve the algorithm robustness, the accuracy of the 3D reconstruction and to provide a richer scene representation. Many consumer products like smartphones, game consoles, augmented and virtual reality devices, as well as cars and household robots are equipped with an increasing amount of cameras and depth sensors. Computer vision systems can highly benefit from this trend by leveraging multiple data sources and providing richer and more accurate results. In this paper, we address the problem of multi-sensor depth map fusion for semantic 3D reconstruction.



Inputs [6] (Std. TSDF Fusion) Learned Fusion (ours)

Figure 1. **Depth map fusion of Kinect and photometric stereo.** Our fusion approach learns sensor noise and outlier statistics and accounts them via confidence weights in the fusion process. This yields more accurate and more complete semantic reconstructions.

Nowadays, depth can be estimated very robustly from multiple and even single RGB images [43]. Nevertheless, depending on the camera, scene lighting, as well as the object and material properties, the noise statistics of computed depth maps can vary largely. Moreover, popular depth sensors like the Kinect have varying noise statistics [54] depending on the depth value and the pixel distance to the image center. They also have trouble recovering depth on object edges as well as on light reflecting or absorbing surfaces, but perform well on low-textured surfaces and within short depth ranges. In contrast, image-based stereo methods usually perform well on object edges and across a wide depth range, but fail on low-textured surfaces and have comparably high noise and outlier rates.

While traditional methods have tried to model these effects, they usually impose strong assumptions about noise distribution, or they require tedious calibration to estimate all parameters [54]. In contrast, we leverage the strength of machine learning techniques to extract sensor properties and scene parameters automatically from training data and use them in form of confidence values for a more accurate semantic depth map fusion. Fig. 1 shows example output of our method. In sum, we make the following **contributions**:

- We propose the first method to unify semantic 3D reconstruction, scene completion and multi-sensor data fusion into a single machine-learning-based framework. Our approach uses only few model parameters and thus needs only small amounts of training data to generalize well.
- Our method analyses the sensor output and learns depth sensor-specific noise and outlier statistics which are considered when estimating confidence values for the TSDF fusion. For the case that the depth source is an algorithm we feed in both information about the depth output and information about the input patches such that out network is better able to learn when the algorithm typically fails.
- Besides the multi-sensor data fusion, our approach can also be used as an expert system for multi-algorithm depth fusion in which the outputs of various stereo methods are fused to reach a better reconstruction accuracy.

2. Related Work

Volumetric Depth Fusion. In their pioneering work, Curless and Levoy [9] proposed a simple and effective method to fuse depth maps from multiple views by averaging *truncated signed distance functions* (TSDFs) within a regular voxel grid. With the broad availability of low-cost depth sensors like the MS Kinect, this method became very popular with influential works like KinectFusion [23] and its numerous extensions, like voxel hashing [36] or voxel octrees [42]. This depth fusion method has become standard for SLAM frameworks like InfiniTAM [24] and was further generalized to account for drift and calibration errors, e.g. ElasticFusion [51], BundleFusion [13], but also for 3D reconstruction frameworks [53, 29, 20, 21, 12, 6].

All these methods have in common that TSDF fusion is performed via simple uniformly weighted averaging. Hence these methods do not account for the fact that depth measurements may exhibit different noise and outlier rates. This has been tackled by probabilistic fusion methods.

Probabilistic Depth Fusion. Probabilistic approaches explicitly model sensor noise, typically with a Gaussian distribution. A very simple approach with only 2.5D output and a Gaussian noise assumption can be found in [16]. A point-based fusion approach is proposed in [25]. Instead of a voxel grid, the fusion updates are directly performed on a point cloud. This has been extended to anisotropic point-based fusion in [34] to account for different noise levels when a surface is observed from a different viewing angle. For a fixed-topology the mesh-based fusion approach by [56] fuses depth information over various mesh resolutions. A more complex probabilistic fusion method is proposed in [52] which includes long range visibility constraint in their online fusion method. A similar model with long-range ray-based visibility constraints was used in [47, 46], although these methods are not real-time capable. Recently, PSDF Fusion [15] demonstrated a combination of probabilistic modeling and a TSDF scene representation. However, they also assume a Gaussian error distribution of the input depth values. Overall, probabilistic approaches handle noise and outliers better than traditional TSDF fusion methods. Nevertheless, the majority of these methods impose strong assumptions about the sensor error distributions to define the prior model. The first method that implicitly learns an unknown error distribution during the fusion is OctNetFusion by Riegler *et al*. [38]. They jointly learn the splitting of the octree scene representation, but multiple sensors or semantic information are not considered.

Multi-Sensor Data Fusion. Early approaches like Zhu *et al*. [55] fuse time-of-flight depth and stereo, but only for a 2.5D depth map. Kim *et al*. [27] fuse the same sensor combination with 3D via a probabilistic framework on a voxel grid. Work by [7] strives for low-level data fusion to improve the Kinect output with stereo correspondences. As an extension of [16], Duan *et al*. [17] use a probabilistic approach for the fusion of Kinect and Stereo in real-time. None of the current multi-sensor depth fusion networks is able to incorporate semantic information and their generalization is usually non-trivial.

3D Reconstruction with Confidences. A wide range of 3D reconstruction approaches estimate confidence values for depth hypotheses which are then later used for adaptive fusion. All these approaches typically use either handcrafted confidence weights [18, 48, 30] rather than learning them intrinsically from data or they learn only 2D score map without learning their 3D fusion [37, 45, 44, 50].

Semantic 3D Reconstruction and Scene Completion. Joint semantic label estimation and 3D geometry has been proposed with traditional energy-based methods to estimate depth maps [32] or dense volumetric 3D [26, 20, 21, 1, 5, 31]. Machine learning-based approaches have pushed the state of the art in reconstructing and completing 3D scenes [10, 14, 12, 6]. These methods are not real-time capable, but real-time fusion of CNN-based single-image depth and semantics has recently been presented in CNN-SLAM [43].

So far none of the semantic 3D reconstruction approaches is able to properly handle multiple sensors with different noise characteristics and their extension is not straightforward. Our goal is a general framework which unifies all the previously discussed properties within a learning-based method.

3. Method

For performing semantic 3D reconstruction, our method requires as input a set of RGB-D images and their corresponding 2D semantic segmentations as shown in Fig. 1. The semantic segmentations can be fused into the TSDF representation of the scene, using [20]. In the following, we

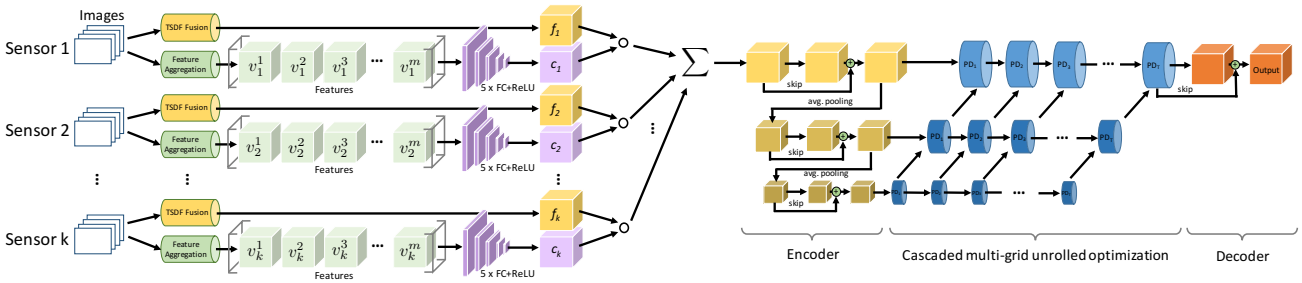


Figure 2. **Network architecture.** Our network consists of two connected networks which are jointly trained. **Left: Sensor Confidence Network** which aggregates voxel-wise confidence values for each sensor. First we fuse semantic TSDFs (yellow) and aggregate features (green) from the input depth maps and images. Then a small fully connected network with ReLU activations processes the features and predicts a confidence weight (magenta). **Right: Semantic 3D Reconstruction Network** which performs 3D reconstruction, denoising and scene completion. This network consists of special layers (blue) which minimize an energy that denoises and completes the scene within a multi-grid setting and finally outputs semantically labeled occupancy grid (red). The right network part corresponds to the one in [6].

describe how we can robustly produce an accurate TSDF by fusing measurements from multiple depth sensors.

Key idea. We consider multiple depth sensors which produce a set of depth maps by scanning a scene. The most common approach to data fusion consists in fusing all the depth maps, regardless of the sensor that produced them, into a TSDF representation of the scene. However, this does not reflect the specific noise and outliers statistics of each measurement. We propose to overcome this issue by learning a confidence estimator for every sensor that weights the measurements before fusing them. For each sensor, we can produce a TSDF representation of the scene by fusing the corresponding depth maps. Our method learns to estimate confidence values for every voxel in TSDF, such that the accuracy of the semantic 3D reconstruction is maximized.

We propose an end-to-end trainable neural network architecture which can be roughly separated into two parts: a *sensor confidence network* which predicts a confidence value for each sensor measurement, and a *semantic 3D reconstruction network* which takes all aggregated noisy measurements and corresponding confidences and performs semantic 3D reconstruction.

The overall network structure is depicted in Fig. 2 and the individual network parts are detailed in the following subsections.

3.1. Sensor Confidence Network

Weighted TSDF Fusion. A sensor i produces a set of depth maps that can be fused into a TSDF f_i, following [9]. We learn to estimate corresponding confidence maps c_i, where for every voxel $\mathbf{x}$, $c_i(\mathbf{x})$ is the confidence for the measurement $f_i(\mathbf{x})$. The fusion of all the sensor measurements is then computed via a point-wise weighted average:

$$f(\mathbf{x}) = \frac{\sum c_i(\mathbf{x}) f_i(\mathbf{x})}{\sum c_i(\mathbf{x})} \tag{1}$$

Goal. The purpose of the confidence weight learning for multi-sensor TSDF fusion is twofold: **1) Intra-Sensor Weighting:** The network captures the noise and outlier statistics among measurements thus producing a spatially varying confidence map, *e.g.* points that are mostly observed from a far distance can get a lower confidence than those mainly observed from a closer distance. **2) Inter-Sensor Weighting:** The network analyses the noise and outlier statistics among different sensors in order to weight them against each other. In this regard the network also accounts for normalization which is important if there are different amounts of data available from different sensors. This avoids for instance a bias towards a sensor with a higher frame rate.

Feature extraction. We aggregate features from the input data which we believe will help the network to estimate a reliable confidence value. Ideally, we could feed all input data into our confidence network and the network could identify important features for the confidence estimation on its own, but the amount of input data for the scenes we consider in this paper renders this infeasible. Therefore, our selected feature set is certainly not exhaustive and there might be other useful features or better feature combinations. However, we found all of them improving the reconstruction results. For each sensor k and each voxel, we extract the following $m = 13$ features $\{v_k^l\}_{l=1}^m$:

- **Average 3×3 patches in depth image** (9 values): Analyzing neighboring depth values helps to identify outliers in the depth map (Fig. 3).
- **Mean and standard deviation of image gradient norm on 3×3 patches** (2 values): Especially for stereo methods the average gradient norm of a patch indicates how much gradient information is contained in the patch. Homogeneously colored patches should lead to low confidence values.
- **Mean and standard deviation of normalized cross correlation (NCC) of stereo 5×5 patches (in case of stereo**

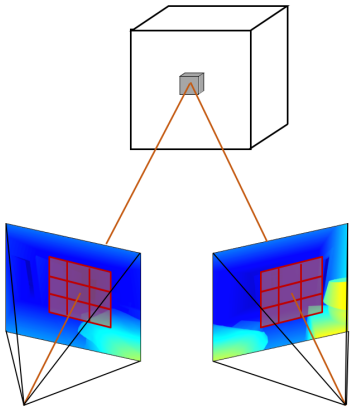


Figure 3. **Voxel-wise feature extraction.** Extraction of the average depth on a 3x3 patch. The voxel is back-projected onto the depth maps, and the patches are centered around the back-projections (represented in red). This approach is identical for the extraction of the gradient and patch similarity features.

algorithms) (2 values): NCC is an established measure for estimating patch similarity for stereo methods. If the patches do not match well, or there is a high variance of NCC values among patches voting for the same point, then the confidence value should be reduced.

This set of features is then processed for each voxel individually by a small neural network which estimates a confidence weight for a single voxel (magenta in Fig. 2).

Confidence Network Architecture. The small confidence estimation networks have identical structure for each sensor and identical weights for each voxel of a sensor. They consist of 5 fully connected neural layers with ReLU activations and with a decreasing number of neurons $\{100, 50, 20, 10, 1\}$. The last layer is initialized with biases equal to one such that the initial confidence values are equal for each sensor. The remaining weights are initialized randomly. The output of the confidence networks are then aggregated into a single TDSF volume which serves as input for a semantic 3D reconstruction network.

3.2. Semantic 3D Reconstruction Network

Our approach learns in an end-to-end fashion how to jointly perform data fusion and semantic 3D reconstruction. The data fusion should facilitate the semantic 3D reconstruction by providing additional and more complete information about the scene. To perform the reconstruction, we use the architecture introduced in [6] which leverages the benefits of neural networks and variational methods. The fundamental principle of the method is to compute a consistent voxel labeling from noisy and incomplete depth such that semantic voxel transitions are statistically similar to the transitions previously seen in the training data. For instance, a bed should be standing on the ground, with vertical transitions to the ground below and the free space above, while a wall should have a horizontal transitions to free space.

The motivations are the following:

- The architecture, which relies on the principles of total variation segmentation and inpainting, contains very few parameters to learn due to weight sharing. Due to the few parameters the network does not need much training data which is beneficial since only few and small real data sets are available for training.
- The compact architecture allows to easily extend the network to estimate further parameters for the data fusion and still allows to process larger scenes with more then 15M voxels.
- The energy formulation allows us to incorporate an arbitrary number of sensors into the 3D reconstruction method, which is more difficult with standard feed-forward architectures.

Variational method. We briefly describe the working principles of the reconstruction network. More details can be found in [6]. At its core, the network minimizes an energy such that the solution corresponds to a scene with label transition statistics that match the training data. We define Ω the voxel grid, and write the energy as:

$$\begin{aligned} &\underset{u}{\text{minimize}} \quad \int_\Omega \left(\|Wu\|_2 + \sum_{s\in\mathcal{S}} (c_s \circ f_s)\, u \right) d\mathbf{x} \qquad (2) \\ &\text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum_{\ell\in\mathcal{L}} u_\ell(\mathbf{x}) = 1 \end{aligned}$$

In Eq. (2), u is the voxel labeling we optimize for, defined such that $u_\ell(\mathbf{x}) \in [0,1]$ is the probability that label ℓ is given to voxel $\mathbf{x}$. The operator $\circ$ denotes element-wise multiplication (Hadamard product). The operator W is a regularizer that enforces the labeling to respect certain conditions on the semantic transitions (*e.g.* the bed stands on the ground). During training, W is learned to capture typical scene statistics. This can be implemented as a convolution which locally compares voxels to their neighborhood, thus verifying the semantic transitions.

The energy (2) is numerically minimized with a first-order algorithm [3]. To this end, dual variables ξ, ν are introduced to account for the non-differentiability and the constraint in Eq. (2), leading to the following equivalent discretized saddle point energy

$$\underset{u}{\text{minimize}} \max_{\|\xi\|_\infty \le 1} \langle Wu, \xi \rangle + \sum_{s\in\mathcal{S}} \langle c_s \circ f_s, u \rangle + \nu \Big(1 - \sum_{\ell\in\mathcal{L}} u_\ell \Big) \qquad (3)$$

The numerical minimization iterations are unrolled and each layer of our network (blue cylinders in Fig. 2) performs the following updates to minimize energy (3). The

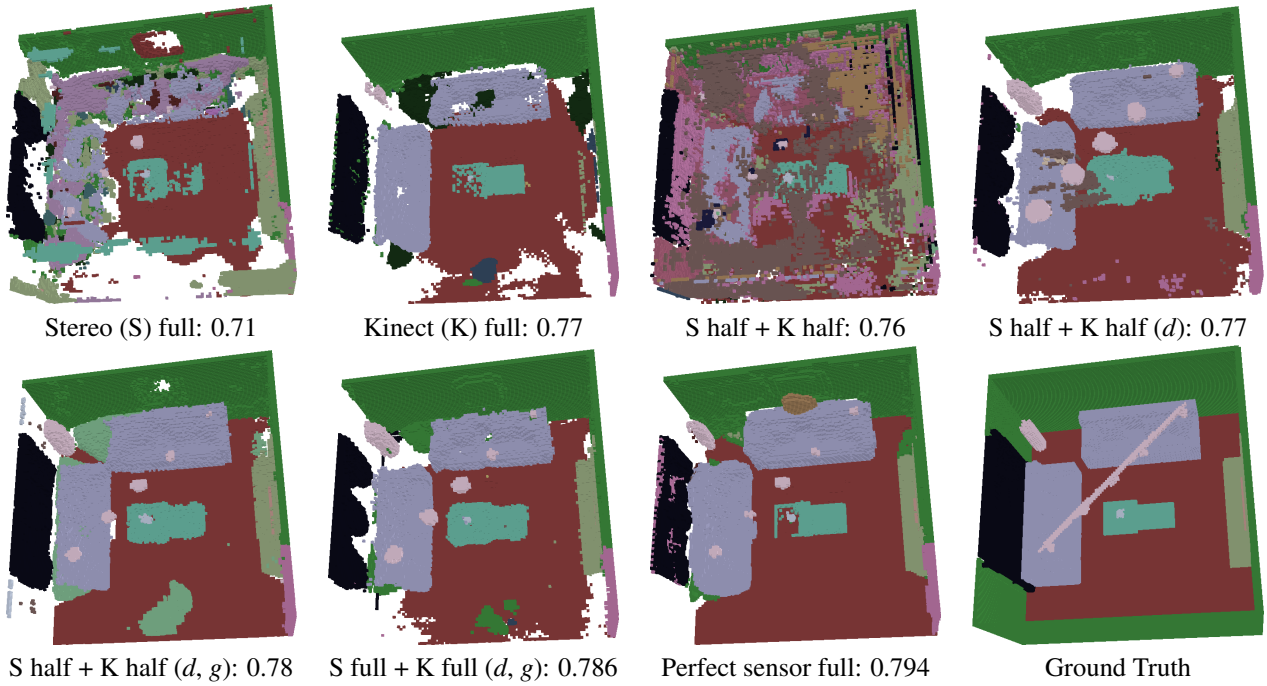


Figure 4. **Ablation study on SUNCG.** For each step we report average semantic accuracy on the whole dataset. Average depth patches for confidence estimation are marked by d, and gradient mean and standard deviation by g. "Full" means that all views were used, whereas "half" means that views were split in two parts. Former refers to noise canceling, and latter to scene completion.

inputs and outputs of each layer are shown on the left.

$$\begin{cases} \nu^{t+1} & = \nu^t + \sigma\left(\sum_\ell \bar{u}_\ell^t - 1\right) \\ \xi^{t+1} & = \Pi_{\|\cdot\|\leq 1}\left[\xi^t + \sigma W \bar{u}^t\right] \\ u^{t+1} & = \Pi_{[0,1]}\Big[u^t - \tau(W^*\xi^{t+1} + \\ & \qquad \nu^{t+1} + \textstyle\sum_{s\in\mathcal{S}} c_s \circ f_s)\Big] \\ \bar{u}^{t+1} & = 2u^{t+1} - u^t \end{cases} \tag{4}$$

For better readability these steps show the single resolution variant. For the multi-grid version the update steps for ξ and u change slightly (please see [6] for more details).

4. Experiments

Setup and Implementation. The entire framework has been implemented in Python/Tensorflow and runs on a computer with E5-2630 processor and an NVidia GTX 1080 Ti GPU running a recent Linux distribution. The network training was done with the ADAM optimizer [28], with learning rate 0.0001 and batch size 4. All training samples were random crops of the input data of dimension $(24, 24, 24)$. Then every crop was randomly rotated around the z-axis and randomly flipped along x and y axes. The network was trained for 1000 epochs, which was enough to converge for all datasets. One epoch iterated once over all scenes. The number of hierarchical levels was set to 3 and number of unrolled optimization iterations to 50, as in [6]. On average training required about 3 hours for 1000 epochs. Inference of one scene takes 3 to 5 minutes on GPU.

Datasets. The experiments were done on three datasets: SUNCG [41], ScanNet [11] and ETH3D [40]. For every dataset and experiment we measure semantic and free-space accuracy. Semantic accuracy (SA) is defined as a ratio of occupied voxels (*i.e.* non free space) for which the particular semantic label was estimated correctly, divided by the total number of occupied voxels. Similarly, the free-space accuracy (FA) is a ratio of voxels, for which the unique free-space label was estimated correctly, divided by the number of free-space voxels. Splitting accuracy into two parts helps to account for domination of free-space voxels in all scenes. Then, the loss function is defined as categorical cross entropy, separately computed for semantic voxels as L_s and free-space voxels as L_f, which are then added together to compute the total loss $L = L_s + \lambda_f L_f$. We set $\lambda_f = 1.5$ to achieve better semantic reconstructions.

4.1. SUNCG

The artificial data origin of the SUNCG dataset with 38 semantic labels enables full control of the data fusion. All components of the method are examined on this dataset with an ablation study. We simulate several different depth sensors, such as a perfect sensor, Kinect and different stereo algorithms.

The baseline is a recent work by Cherabier *et al.* [6]

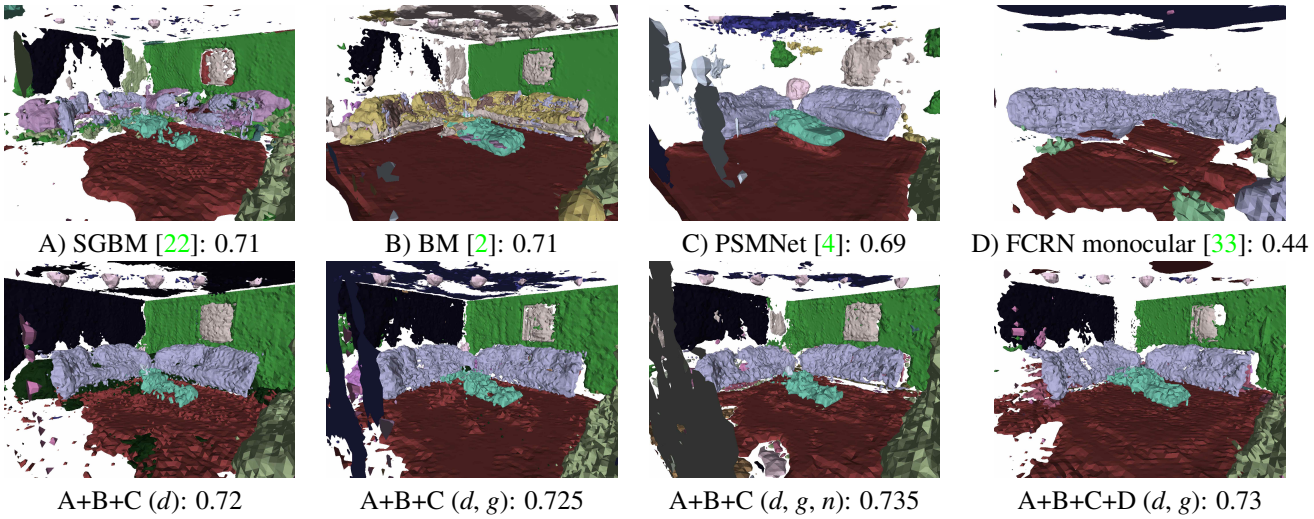


Figure 5. **Close-up views for the expert system experiment with learned fusion of 4 stereo algorithms**: Block Matching (BM) [2], Semi-Global Block Matching (SGBM) [22], PSMNet [4], FCRN monocular [33]. The top row contains TSDF fusion results from each of the stereo algorithms separately, whereas the bottom row provides results of different algorithm combination as well as different types of input used for confidence estimation. Average depth patches for confidence estimation are denoted by d, gradient mean and standard deviation by g, and average normalized cross correlation of stereo patches by n. FCRN is a monocular method and thus n cannot be computed. For each method we report average semantic accuracy on the whole dataset.

where they use a simple averaging of input TSDF volumes trained with the network without confidence estimation module. We add gradually the following input measurements: average 3×3 depth patches, mean and standard deviation of gradients, mean and standard deviation of normalized cross-correlation between stereo patches in case of a stereo algorithm.

Training and validation sets were created by randomly selecting 100 and 30 scenes respectively. Qualitative results for a selected scene and quantitative results on the whole dataset are shown in Fig. 4. Every input brings an increase in performance, measured by semantic and free-space accuracy. Quantitative results contain only semantic accuracy. The free-space accuracy was close to 0.95 with small deviations in all settings. The increase in accuracy is small, but the values approach the upper limit given by the perfect sensor and the reconstructions look better visually.

4.2. Stereo Expert System

The proposed method was applied to create an expert system for stereo algorithms. We used the following four methods for stereo depth estimation:

- Pyramid Stereo Matching Network (PSMNet) [4] – 3D CNN architecture with spatial pyramid pooling module for depth map estimation from a stereo pair.
- Depth Prediction with Fully Convolutional Residual Networks (FCRN) [33] – fully convolutional architecture with residual learning which is trained to estimate depth map from a single RGB image.

Method	TP rate	Distance	SA	FA
Input	0.507	3.376	0.55	0.79
ScanComplete [14]	0.588	2.527	0.47	0.90
Standard TSDF in [6]	0.837	1.606	0.79	0.96
Proposed	**0.953**	**1.410**	**0.90**	**0.97**

Table 1. **Quantitative results on the ScanNet dataset [11].** We report true positive (TP) rate of completion, average surface distance to the ground truth, semantic accuracy (SA) and free-space accuracy (FA). The comparison demonstrates that our method performs better than the baselines [6, 14].

- Semi-Global Block Matching (SGBM) [22] – classical method (by H. Hirschmuller), which matches blocks of a given size in a pair of images using mutual information.
- Block Matching (BM) [2] – a version of block matching algorithm provided by K. Konolige.

At first, we trained a network without confidence values on each of the stereo algorithms separately. Then, a fused combination of these methods with learned confidence values was trained. Fig. 5 shows that the learned fusion performs better than any of the stereo methods on its own. More importantly, the learned fusion results are less noisy, more accurate and complete. The stereo system results can be compared to results of other sensor fusion models in Fig. 4.

4.3. ScanNet

Previous two experiments were done on a synthetic dataset. The next evaluation on ScanNet [11] dataset shows that the method is able to perform well also on real data. However, the dataset contains only measurements from one

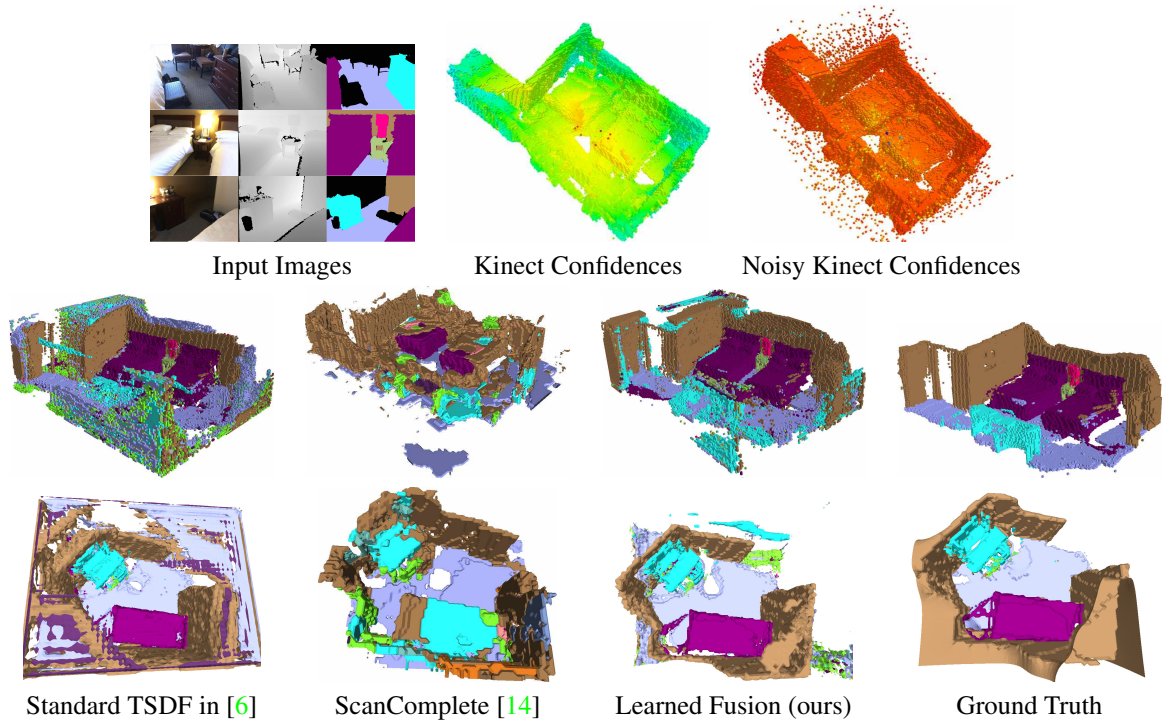


Figure 6. **Our learned fusion approach in comparison to standard TSDF averaging in [6].** We fuse real world Kinect data with an artificially noised Kinect sensor. Our approach leads to less artificats, more consistent semantic labels and better accuracy scores in comparison to the ground truth data. **Top row**: Estimated confidence values from input measurement on ScanNet dataset. **Middle and bottom rows**: The proposed learned fusion in comparison to datacost averaging in [6], ScanComplete [14] and ground truth.

Dataset	Fusion Method	SA	FA
ETH3D	Standard TSDF in [6]	0.50	0.96
	Learned (proposed)	**0.59**	**0.97**

Table 2. **Evaluation on the ETH3D dataset [40].** We report semantic accuracy (SA) and free-space accuracy (FA). The proposed learned fusion outperforms the baseline [6] with standard TSDF fusion, when used with different sensors.

sensor, Kinect. In order to create an additional sensor, we simulated an artificial noisy Kinect with outliers modeled by Gaussian noise with zero mean and standard deviation of 2 meters with probability of 1%. We used 7 training scenes and 5 validation scenes from the hotel bedroom category, which have 9 semantic labels. Ground truth was obtained by running total variation on all views, whereas only every 10th view was used for further fusion. The proposed fusion method was compared to two state-of-the-art baselines, Cherabier *et al*. [6] with simple averaging of TSDF volumes and ScanComplete [14]. ScanComplete also optimizes geometry, but is not designed for sensor fusion. Hence, this method performs worse when we input uniformly averaged multi-sensor data. ScanComplete is trained on SUNCG and fine-tuning on ScanNet is difficult due to incompleteness as already stated by the authors and thus omitted.

Tab. 1 shows various performance scores of the input geometry in comparison to the completed results. The proposed learned fusion improves semantic accuracy of [6] by 11%. Volumes with estimated confidence values are visualized in Fig. 6, together with two selected reconstructed validation scenes. Learned confidence values in the top row show that the network is able to learn different weights for artificially created noisy Kinect sensor and downscales occasional noisy pixels. Voxels outside walls are not downscaled and not penalized, because they are part of the *unknown* label which is not included in the loss function. For the non-noisy Kinect, the confidence values are decreasing for voxels further away from the center. The Kinect sensor is known to have less precise measurements with increasing depth [54] and this was learned by the network.

4.4. ETH3D

The last experiment on ETH3D dataset is done to confirm that the proposed method is able to work not only on real data, but also with several real sensors. ETH3D dataset [40] comprises multi-view images with high resolution camera, as well as with low resolution camera rigs. The ground truth is given by a laser scan.

We again tested that the joint learned fusion performs

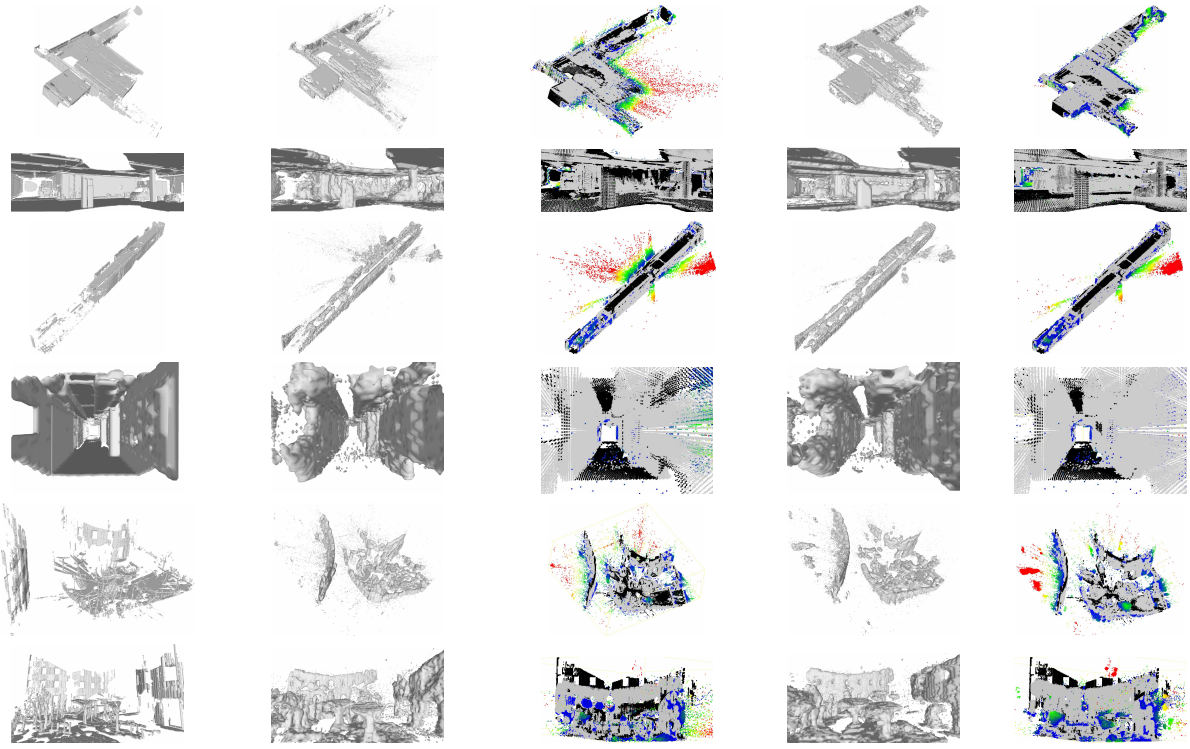


Figure 7. **Experiment on ETH3D**. Two top rows contain the first training scene – *delivery area*, and two middle rows another training scene – *terrains*. Two bottom rows contain the validation scene – *playground*. For each scene we show a full view and a close-up view. The distance to the ground truth (error) is color-coded by a gray color where the distance is less than 5 voxels (correct reconstruction) and blue-green-yellow-red color scale for outliers. Black voxels denote regions where the ground truth geometry was not reconstructed.

better than [6] with simple TSDF averaging. Training set had two scenes, *delivery area* and *terrains*, whereas validation set consisted of a single scene *playground*. Only these three scenes contain measurements from both sensors, which explains the used set size. The resolution was set to 8 cm, which gives scenes large enough for training. The number of parameters to learn is low and the results show that only several scenes are enough to train the model. The label set consists of only two labels, *free-space* and *occupied-space*, as no semantic ground truth is available.

Tab. 2 contains quantitative results on ETH3D dataset, where the increase in semantic accuracy is 9%. Fig. 7 shows visualized reconstructions of all scenes with one close-up view for each scene. The learned fusion is able to provide more complete reconstructions and it does not contain as many separate outlier semantic voxels in ground truth free-space. The error is measured as distance to the ground truth with gray color regions representing the correct reconstruction with error less than 5 voxels.

5. Conclusion

We proposed a novel machine learning-based depth fusion method that unifies semantic depth integration, multi-sensor or multi-algorithm data fusion as well as geometry denoising and completion. We substantially generalize the recent semantic 3D reconstruction method [6] to incorporate an arbitrary amount of depth sensors. To balance the contribution of each sensor according to their noise statistics, we extract features from the sensor data and learn the network to predict suitable confidence weights for each sensor and each point in space. Our approach is generic and can also learn reliability statistics of different stereo algorithms. This allows us to use the method as an expert system that weights and fuses the outputs of all algorithms, providing a result that is better than of any individual algorithm.

Acknowledgements. Denys Rozumnyi was supported by Czech Science Foundation grant GA18-05360S, CTU student grant SGS17/185/OHK3/3T/13 and ETH SSRF. Further support was received by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00280. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Maroš Bláha, Christoph Vogel, Audrey Richard, Jan D. Wegner, Thomas Pock, and Konrad Schindler. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[3] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, May 2011. 4

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 6

[5] Ian Cherabier, Christian Häne, Martin R. Oswald, and Marc Pollefeys. Multi-label semantic 3d reconstruction using voxel blocks. In *International Conference on 3D Vision (3DV)*, 2016. 1, 2

[6] Ian Cherabier, Johannes L. Schönberger, Martin R. Oswald, Marc Pollefeys, and Andreas Geiger. Learning priors for semantic 3d reconstruction. In *European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 3, 4, 5, 6, 7, 8

[7] Walon Wei-Chen Chiu, Ulf Blanke, and Mario Fritz. Improving the kinect by cross-modal stereo. In *Proc. of the British Machine and Vision Conference (BMVC)*, pages 1–10, 2011. 2

[8] Yan Cui, Sebastian Schuon, Derek Chan, Sebastian Thrun, and Christian Theobalt. 3d shape scanning with a time-of-flight camera. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1180, 2010. 1

[9] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312, 1996. 2, 3

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5, 6

[12] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 458–474, 2018. 1, 2

[13] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3):24:1–24:18, 2017. 2

[14] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 6, 7

[15] Wei Dong, Qiuyuan Wang, Xin Wang, and Hongbin Zha. Psdf fusion: Probabilistic signed distance function for on-the-fly 3d data fusion and scene reconstruction. In *Proc. European Conference on Computer Vision (ECCV)*, September 2018. 2

[16] Yong Duan, Mingtao Pei, and Yunde Jia. Probabilistic depth map fusion for real-time multi-view stereo. In *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, pages 368–371, 2012. 2

[17] Yong Duan, Mingtao Pei, and Yucheng Wang. Probabilistic depth map fusion of kinect and stereo in real-time. In *2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012, Guangzhou, China, December 11-14, 2012*, pages 2317–2322, 2012. 2

[18] Simon Fuhrmann and Michael Goesele. Floating scale surface reconstruction. *ACM Trans. Graph.*, 33(4):46:1–46:11, 2014. 2

[19] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8):1362–1376, 2010. 1

[20] Christian Häne, Christopher Zach, Andrea Cohen, Roland Angst, and Marc Pollefeys. Joint 3d scene reconstruction and class segmentation. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104, 2013. 1, 2

[21] Christian Häne, Christopher Zach, Andrea Cohen, and Marc Pollefeys. Dense semantic 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1730–1743, 2017. 1, 2

[22] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb 2008. 6

[23] Shahram Izadi, Richard A. Newcombe, David Kim, Otmar Hilliges, David Molyneaux, Steve Hodges, Pushmeet Kohli, Jamie Shotton, Andrew J. Davison, and Andrew W. Fitzgibbon. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2011, Vancouver, BC, Canada, August 7-11, 2011, Talks Proceedings*, page 23, 2011. 1, 2

[24] Olaf Kähler, Victor Adrian Prisacariu, Carl Yuheng Ren, Xin Sun, Philip H. S. Torr, and David W. Murray. Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Trans. Vis. Comput. Graph.*, 21(11):1241–1250, 2015. 2

[25] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion.

In *2013 International Conference on 3D Vision, 3DV 2013, Seattle, Washington, USA, June 29 - July 1, 2013*, pages 1–8, 2013. 2

[26] Byung-Soo Kim, Pushmeet Kohli, and Silvio Savarese. 3d scene understanding by Voxel-CRF. In *Proc. International Conference on Computer Vision (ICCV)*, pages 1425–1432, 2013. 1, 2

[27] Y.M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Micusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *IEEE Workshop on 3-D Digital Imaging and Modeling (3DIM) at the International Conference on Computer Vision (ICCV)*, 2009. 2

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, abs/1412.6980, 2014. 5

[29] Kalin Kolev, Maria Klodt, Thomas Brox, and Daniel Cremers. Continuous global optimization in multiview 3d reconstruction. *International Journal of Computer Vision*, 84(1):80–96, 2009. 2

[30] Andreas Kuhn, Heiko Hirschmüller, Daniel Scharstein, and Helmut Mayer. A TV prior for high-quality scalable multi-view stereo reconstruction. *International Journal of Computer Vision*, 124(1):2–17, 2017. 2

[31] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M. Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *Proc. European Conference on Computer Vision (ECCV)*, pages 703–718. Springer, 2014. 1, 2

[32] Lubor Ladický, Paul Sturgess, Chris Russell, Sunando Sengupta, Yalin Bastanlar, William Clocksin, and Philip H. S. Torr. Joint Optimization for Object Class Segmentation and Dense Stereo Reconstruction. *International Journal of Computer Vision*, 100(2):122–133, November 2012. 2

[33] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 6

[34] Damien Lefloch, Tim Weyrich, and Andreas Kolb. Anisotropic point-based fusion. In *18th International Conference on Information Fusion, FUSION 2015, Washington, DC, USA, July 6-9, 2015*, pages 2121–2128, 2015. 2

[35] Parsa Mirdehghan, Wenzheng Chen, and Kiriakos N. Kutulakos. Optimal structured light à la carte. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[36] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6):169:1–169:11, 2013. 2

[37] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *Proc. of the British Machine and Vision Conference (BMVC)*, 2016. 2

[38] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *International Conference on 3D Vision (3DV)*, 2017. 2

[39] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. European Conference on Computer Vision (ECCV)*, 2016. 1

[40] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 7

[41] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[42] Frank Steinbrücker, Christian Kerl, and Daniel Cremers. Large-scale multi-resolution surface reconstruction from RGB-D sequences. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 3264–3271, 2013. 2

[43] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6565–6574, 2017. 1, 2

[44] Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *Proc. European Conference on Computer Vision (ECCV)*, pages 323–338, 2018. 2

[45] Fabio Tosi, Matteo Poggi, Stefano Mattoccia, Alessio Tonioni, and Luigi di Stefano. Learning confidence measures in the wild. In *Proc. of the British Machine and Vision Conference (BMVC)*, 2017. 2

[46] Ali Osman Ulusoy, Michael J. Black, and Andreas Geiger. Patches, planes and probabilities: A non-local prior for volumetric 3d reconstruction. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3280–3289, 2016. 2

[47] Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *2015 International Conference on 3D Vision, 3DV 2015, Lyon, France, October 19-22, 2015*, pages 10–18, 2015. 2

[48] Benjamin Ummenhofer and Thomas Brox. Global, dense multiscale reconstruction for a billion points. *International Journal of Computer Vision*, 125(1-3):82–94, 2017. 2

[49] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(5):889–901, 2012. 1

[50] Chamara Saroj Weerasekera, Thanuja Dharmasiri, Ravi Garg, Tom Drummond, and Ian D. Reid. Just-in-time reconstruction: Inpainting sparse maps using single view depth predictors as priors. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1–9, 2018. 2

[51] Thomas Whelan, Renato F. Salas-Moreno, Ben Glocker, Andrew J. Davison, and Stefan Leutenegger. Elasticfusion:

Real-time dense SLAM and light source estimation. *I. J. Robotics Res.*, 35(14):1697–1716, 2016. 2

[52] Oliver J. Woodford and George Vogiatzis. A generative model for online depth fusion. In *Proc. European Conference on Computer Vision (ECCV)*, pages 144–157, 2012. 2

[53] Christopher Zach, Thomas Pock, and Horst Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In *Proc. International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 2

[54] Bernhard Zeisl and Marc Pollefeys. Structure-based autocalibration of RGB-D sensors. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 5076–5083, 2016. 1, 7

[55] Jiejie Zhu, Liang Wang, Ruigang Yang, and James Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[56] Jacek Zienkiewicz, Akis Tsiotsios, Andrew J. Davison, and Stefan Leutenegger. Monocular, real-time surface reconstruction using dynamic level of detail. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 37–46, 2016. 2