Stereo and ToF Data Fusion by Learning from Synthetic Data

Gianluca Agresti[1], Ludovico Minto[1], Giulio Marin[1], Pietro Zanuttigh[1]

University of Padova, Via Gradenigo 6B, 35131 Padova, Italy

Abstract

Time-of-Flight (ToF) sensors and stereo vision systems are both capable of acquiring depth information but they have complementary characteristics and issues. A more accurate representation of the scene geometry can be obtained by fusing the two depth sources. In this paper we present a novel framework for data fusion where the contribution of the two depth sources is controlled by confidence measures that are jointly estimated using a Convolutional Neural Network. The two depth sources are fused enforcing the local consistency of depth data, taking into account the estimated confidence information. The deep network is trained using a synthetic dataset and we show how the classifier is able to generalize to different data, obtaining reliable estimations not only on synthetic data but also on real world scenes. Experimental results show that the proposed approach increases the accuracy of the depth estimation on both synthetic and real data and that it is able to outperform state-of-the-art methods.

Keywords: Sensor Fusion, Stereo, Time-of-Flight, Deep Learning

1. Introduction

There exist many different devices and algorithms for real-time depth estimation including active lighting devices and passive systems using only regular cameras. The first family includes structured light cameras and Time-of-Flight (ToF) sensors while the most notable example of the second family are the stereo cameras. None of these solutions is completely satisfactory.

[1] All authors equally contributed to the work.

Active devices like Time-of-Flight and structured light cameras are able to robustly estimate the 3D geometry independently of the scene content but they have a limited spatial resolution, a high level of noise and a reduced accuracy on low reflective surfaces. Passive stereo vision systems, although widely used for their simplicity and low cost of the hardware setup, have various limitations, in particular their accuracy strongly depends on the scene content and the acquisition is not very reliable on uniform or repetitive regions. On the other side, passive stereo vision systems have a high resolution and a limited amount of noise. The characteristics of the two families of devices are complementary and the fusion of data from the two systems has been the subject of several research studies in the last years.

This paper proposes a depth estimation algorithm that combines stereo and ToF data together extending the approach presented in [1]. Several approaches have been proposed for the fusion of stereo and ToF data (see Section 2), but they all rely on deterministic schemes, while the proposed method is the first to use machine learning (and more specifically deep learning) for this task.

An effective solution for this task needs two fundamental tools: the estimation of the reliability of the data acquired by the two devices at each location and a fusion algorithm that uses this information to properly combine the two data sources. The reliability of ToF data has been traditionally estimated by modeling the noise of such sensors [2]. ToF sensors are typically affected by various sources of error. Shot noise can be estimated from the amplitude and intensity of the received signal, but the estimation of the impact of errors related to ToF working principles, like mixed pixels and the multipath error, is more challenging. In particular, the latter is very difficult to be directly estimated and compensated, because the light rays are scattered multiple times before reaching the sensor. A key contribution of this work is the use of a machine learning framework to estimate confidence information for ToF data. Deep learning techniques and in particular Convolutional Neural Networks (CNNs) increase the performance of many computer vision tasks including the estimation of depth data reliability [3]. However, CNNs have never been applied to the estimation of ToF data reliability, mostly due to the lack of large datasets with ToF depth and ground truth information, their acquisition being challenging and time consuming. For this reason, in this paper we investigate the possibility of training a suitable CNN using synthetic data while testing its performance on real world data. This is a key difference with [1] that was dealing with synthetic data only.

Stereo confidence data are typically estimated with metrics based on the analysis of the shape of the cost function [4]. These metrics capture the effects of the local matching cost computation, but most recent stereo vision techniques use more complex global optimization schemes whose behavior is not captured by standard metrics. To obtain an estimation of the confidence that is more accurate and coherent with the ToF data, we use the same deep learning framework to jointly estimate the stereo and ToF confidences. In data fusion applications, confidence information is used to decide which data source should be trusted more and what really matters is the ratio between the two confidence values at each location rather than their absolute value. By using a single deep network jointly estimating the two measures we obtain a confidence measure that fits particularly well the fusion application.

More in detail, the proposed algorithm extends the work presented in [1]: it starts from reprojecting the ToF data to the stereo camera viewpoint in order to have all the data in the same reference system. Also, ToF data are upsampled to the spatial resolution of the stereo setup by using a combination of segmentation clues and bilateral filtering [5]. Then, confidence information for both ToF and stereo depth data are jointly estimated with an ad-hoc CNN that takes in input multiple clues, i.e., the stereo and ToF disparities, the ToF amplitude and the difference between the reference image and the target one warped over it according to disparity information, providing a hint of the stereo matching accuracy. The construction of the input data for the network and the deep learning architecture include some difference w.r.t. [1] to improve the performance and the generalization capabilities when testing the framework on real world data. Finally, we use an extended version of the Local Consistency (LC) framework [5, 6] that is capable of using the confidence information to perform the fusion of the two data sources.

As already pointed out, CNNs training requires a good amount of data with the corresponding ground truth information. At the time of writing there are no available datasets collecting these data and furthermore the acquisition of accurate ground truth data for real world 3D scenes is a challenging operation. For this reason we rendered 55 different 3D synthetic scenes using *Blender* [7] with examples of various acquisition issues including reflections, global illumination and repetitive patterns. Realistic stereo and ToF data have been simulated for the rendered scenes using *LuxRender* [8] and a simulator realized by Sony EuTEC starting from the work of Meister et al. [9]. We used this dataset, that represents another contribution of this paper, to train the proposed CNN. The use of a larger dataset also allowed us to perform a more reliable experimental evaluation with respect to competing approaches that have been typically tested only on a few sample scenes. To evaluate the effectiveness of our approach in the real world scenario, we also acquired a real world dataset using a Kinect v2 and a ZED stereo camera. Ground truth depth information for this dataset has also been acquired. The proposed confidence estimation strategy not only proved to be able to accurately estimate a confidence measure for both stereo and ToF synthetic depth data but also demonstrated good generalization properties being able to properly estimate the reliability of real world data even if only synthetic information was used in the training process. The results have also been computed on a third dataset used by some previous works [5, 6] to provide a more extensive comparison with state-of-the-art approaches.

The related works are summarized in Section 2. Then, Section 3 introduces the general architecture of the proposed approach. Section 4 describes the deep learning network used to compute confidence information. The fusion algorithm is described in Section 5. The real and synthetic datasets are described in Section 6 and the results are discussed in Section 7. Finally, Section 8 draws the conclusions.

2. Related Works

Depth estimation using stereo vision cameras is a long term research field and a large number of different approaches have been proposed and tested on public data like the Middlebury [10] and KITTI [11] benchmarks. A good review on this topic is [12]. Despite the large amount of research and the continuous improvement of the performance of these methods, the depth estimation accuracy of stereo systems depends on many factors, and in particular on the photometric content of the scene. The estimation is less accurate in regions with fewer details, i.e., when the scene contains a limited amount of texture, or on repetitive patterns. Since the accuracy can vary considerably between different scenes or even different regions of the same scene, it is important to estimate the confidence of the computed data. Until a few years ago, the confidence information for stereo systems used to be computed mostly by analyzing some key properties of the stereo matching cost function. A comprehensive review of this family of approaches is [4]. Recently, machine learning thechniques started to be used for this task, first with traditional approaches (e.g., Random Forests), then by using deep learning techniques. A very recent review of machine learning approaches for stereo confidence computation is [3]. An example of approach of this family is [13], that uses a CNN to estimate the confidence information from image patches. A two channel image patch representation is used also by [14], while [15] improves standard confidence metrics by enforcing the local consistency of the confidence maps with a deep network.

On the other side, ToF cameras represent a quite robust solution for depth acquisition [16, 17, 2, 18, 19, 20]. The various low cost depth cameras available on the market can acquire depth information in real-time and are more robust to the scene content with respect to stereo systems, in particular they can estimate the depth also in regions without texture or with repetitive patterns. On the other side, ToF cameras have their own limitations, e.g., the resolution is typically lower than standard cameras and they are noisy. These cameras are also affected by other sources of errors like the multi-path interference and the mixed pixel effect. A detailed analysis of the various error sources has been presented in [19] while [20] focuses on the effects of the reflectivity of the scene on the depth accuracy. There exist large datasets acquired with ToF sensors for other computer vision applications like semantic segmentation [21, 22], gesture recognition [23] and face recognition [24], but they all lack ground truth depth data that is very time consuming to acquire. For this reason the confidence of ToF data is typically computed with deterministic schemes. A very recent work [25] uses deep learning for ToF data denoising.

ToF cameras and stereo vision systems rely on completely different depth estimation principles. For this reason, they have complementary characteristics and the fusion of the data acquired from the two sources should produce more accurate measures. Several different approaches for the combination of stereo and ToF data have been proposed. Comprehensive reviews of the topic can be found in [26] and [2].

A possible approach is to model the problem with a MAP-MRF Bayesian formulation and optimize a global energy function with belief propagation. This technique has been used by various works of Zhu et Al. [27, 28, 29]. A probabilistic formulation has been used in [30] that computes the depth map with a ML local optimization. The approach has been extended in [31] that adds a global MAP-MRF optimization scheme. A second possibility is to use a a variational fusion framework. Examples of this family are the methods of [32], that also uses confidence measures for the ToF and stereo vision systems to drive the process, and the works of Chen et Al. [33, 34], that combines the variational approach with edge-preserving filtering.

A different solution is proposed in [35], that computes the depth data by solving a set of local energy minimization problems. Another solution is to use a locally consistent framework [36] to fuse the two data sources. The idea has been firstly introduced in [5], then improved in [6] by adding the confidence information for the two data sources. Finally [1], that represent the starting point for the proposed work, extends the work of [6] by introducing a more refined confidence estimation strategy relying on deep learning techniques. However this work deals only with synthetic data while in this journal extension we extend the approach to real world scenes.

Another task related to stereo-ToF data fusion is the improvement of ToF depth with the information coming from a single color camera. For this task many different strategies have been proposed [37, 38, 39, 40, 41, 42]. Common approaches include solutions based on bilateral filtering [38, 39], on edge-preserving interpolation schemes [41] and on the development of confidence information for ToF data [40].

3. Proposed Method

The target of the proposed work is to combine the data from a ToF camera with a stereo vision system in order to extract an accurate depth representation. Both devices are able to produce an estimation of depth data from the corresponding viewpoint and the proposed method combines these two representations to provide a dense and more accurate depth map from the point of view of one of the color cameras of the stereo setup.

The combination of the two depth fields requires to firstly bring the data into a common reference system. To this purpose it is necessary to jointly calibrate the two sensors. In this work the experimental evaluation is performed with both synthetic and real world data. The calibration task is trivial for the case of synthetic information, since camera parameters can be directly extracted from the simulation software, but the accurate calibration of real world Time-of-Flight and stereo systems is challenging.

Depth cameras are usually pre-calibrated by proprietary algorithms, and the calibration parameters are stored in the device during manufacturing and made accessible to the user only by official drivers. The manufacturer calibration allows to extract 3D locations relative to the camera viewpoint but for the fusion application it is also necessary to get the relative position between the depth camera and the stereo setup.

In order to solve this task we used an extension of the approach of Zhang [43] for camera calibration with a regular black and white checkerboard. This method requires to acquire images of the planar checkerboard from different positions and orientations. We collected approximately 30 images of a checkerboard visible from all the sensors and then run a checkerboard corner detector on all the images obtaining for each camera n (in our case there are just 3 cameras) and for each pose k a set of J points $\mathbf{p}_{n,k}^{j}$. For the ToF camera we used the amplitude image (the corners are obviously not visible on the depth data) while for the stereo camera we used the color images. The calibration parameters are estimated by minimizing the Euclidean distance between the planar positions of the measured and the projected 3D points after anti-distortion, given by

$$\min_{\mathbf{K}_n;\mathbf{d}_n;\mathbf{R}_{n,k};\mathbf{t}_{n,k}} \sum_{n=1}^{N}\sum_{k=1}^{M}\sum_{j=1}^{J} \delta_{n,k}^{j} \|\mathbf{p}_{n,k}^{j} - f(\mathbf{K}_n;\mathbf{d}_n;\mathbf{R}_{n,k};\mathbf{t}_{n,k};\mathbf{P}^j)\|_2^2 \qquad (1)$$

where $\mathbf{p}_{n,k}^{j}$ is the projection of the 3D feature P^j with coordinates $\mathbf{P}^j$ on the n-th camera at the k-th pose of the checkerboard, $\delta_{n,k}^{j}$ is 1 if P^j is visible by the n-th camera at the k-th pose and 0 otherwise. The function $f(\mathbf{K}_n;\mathbf{d}_n;\mathbf{R}_{n,k};\mathbf{t}_{n,k};\mathbf{P}^j)$ accounts for projection and distortion. The minimization of (1) is solved by nonlinear optimization techniques such as the Levenberg-Marquardt method. Matrices $\mathbf{R}_{n,k}$ and $\mathbf{t}_{n,k}$ describes the k-th checkerboard pose with respect to the n-th camera. Given that $\mathbf{R}_{n\to m}$ and $\mathbf{t}_{n\to m}$ are the rotation and translation matrices relating cameras n and m, the following relationships hold

$$\begin{aligned} \mathbf{R}_{m,k} &= \mathbf{R}_{n,k}\mathbf{R}_{n\to m} \\ \mathbf{t}_{m,k} &= \mathbf{R}_{n,k}\mathbf{t}_{n\to m} + \mathbf{t}_{n,k}. \end{aligned} \qquad (2)$$

from which one can retrieve the pose of a given camera with respect to the reference camera.

The proposed algorithm is divided into four main steps (see Figure 1):

1. The depth information acquired from the ToF sensor is reprojected to the reference color camera viewpoint and interpolated to the same resolution of the color cameras. The interpolation is necessary since ToF sensors have typically a low resolution, specially if compared with

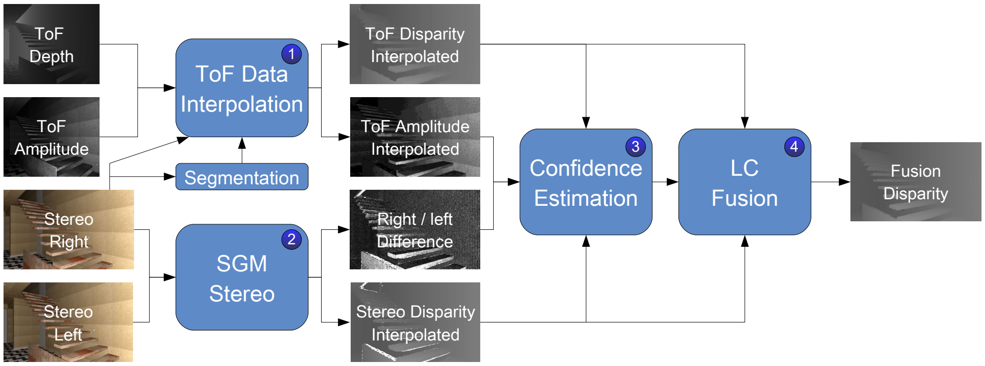


Figure 1: Flowchart of the proposed approach.

modern color cameras. The approach used for this task has been derived from [5]: we used an extended version of the cross bilateral filter where the filter is driven by three terms, the standard spatial Gaussian weighting, the range term computed on the color image and an additional segmentation-based term that depends on a segmented version of the color image computed with Mean-Shift clustering [44]. This procedure allows us to produce a high resolution depth map aligned with the color camera lattice that will be used by the fusion algorithm. Finally, the depth map is converted to a disparity map, since the fusion algorithm works in disparity space. More details on this step can be found in [5, 6].

2. In parallel, the the Semi-Global Matching (SGM) stereo vision algorithm [45] is used to compute a high resolution disparity map from the stereo pair. We selected this algorithm since it provides a good compromise between computation time and performance, however the proposed approach is independent of the selected stereo vision algorithm.
3. After obtaining the two disparity fields, confidence information is jointly estimated for the stereo and ToF disparity maps using the CNN architecture presented in Section 4.
4. Finally the reprojected and interpolated ToF disparity and the stereo disparity are fused using an extended version of the Locally Consistent (LC) algorithm [36, 6]. This step is described in Section 5.

4. Confidence Estimation with Deep Learning

A fundamental step in order to reliably fuse the two disparity maps is their per-pixel confidence estimation. To this purpose, we designed and trained a 6-layer CNN that takes in input different clues from ToF and stereo data and jointly uses the information from both devices to infer the two confidence maps. In particular, the proposed CNN takes in input four channels associated to the following clues:

- A difference map Δ encoding for mismatches between corresponding visual cues in the stereo image pair.
- The stereo disparity map D_S.
- The ToF disparity map D_T obtained from the ToF depth map after reprojection on the reference camera and conversion to the disparity space.
- The ToF amplitude image reprojected on the reference camera of the stereo vision system A_T.

Since raw input data correspond to different sources of information coming from heterogeneous sensors, a lightweight pre-processing stage is needed to convert such data into the desired form.

The first clue Δ aims providing a rough measure of the accuracy of the disparities computed by the stereo algorithm. The idea is that accurate disparity estimates are likely to result in pairs of corresponding pixels with similar values in the reference and target stereo images respectively. On the contrary, corresponding pixels computed using inaccurate disparities are likely to hold different values since they correspond to different parts of the scene. In order to compute Δ, both the reference and target stereo images are first converted to grayscale images giving I_R and I_T respectively. The target grayscale image I_T is then reprojected on the reference camera using the stereo disparity, thus obtaining the image I'_T. Finally, the absolute difference between I_R and I'_T is taken, leading to

$$\Delta = \left| I_R - I'_T \right| \tag{3}$$

The stereo disparity clue D_S is directly obtained from the stereo disparity map while the ToF disparity clue D_T is derived from the ToF depth map first by reprojecting it on the reference viewpoint then by converting it to the disparity space. Similarly, the last clue A is derived by reprojecting the ToF amplitude image onto the reference frame. Finally, the four clues Δ, D_T, D_S and A_T are packed together in a four-channel input tensor where each channel is independently normalized to the unit interval by applying an appropriate scaling factor. Such tensor can be fed to the CNN to produce as output two confidence maps P_T and P_S for the ToF and stereo disparity respectively.

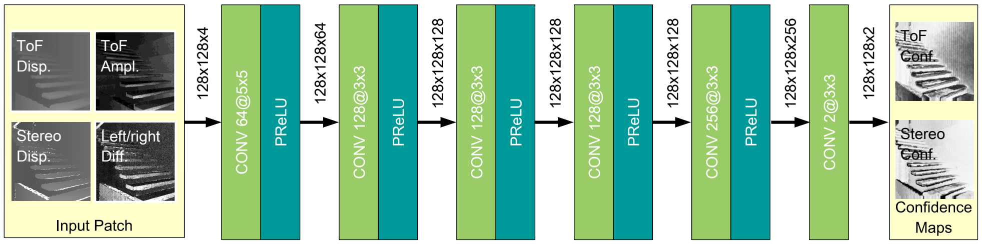


Figure 2: Architecture of the proposed deep learning framework. A 4-channel training patch of size 128×128 [pxl] is fed to a CNN with 6 convolutional layers: the figure shows the number of filters, their spatial kernel sizes and the size of the outputs for each layer.

The proposed CNN architecture is shown in Figure 2. The network is made of a stack of six convolutional layers (CONV) each followed by a Parametric Rectified Linear Unit (PReLU) activation layer, except for the last convolutional layer. The PReLU activation function [46] has been chosen over the standard Rectified Linear Unit (ReLU) activation to prevent the dead-neuron effect caused by negative inputs entering the ReLU zero-slope region. In our experiments, we set the slope of the negative part of the PReLU activation function to 0.02.

The first five layers are assigned an increasing number of filters, namely 64, 128, 128, 128 and 256 respectively. Filter kernels in the first convolutional layer have a spatial size of 5×5 [pxl], while kernels in all subsequent layers are 3×3 [pxl] wide. The last convolutional layer has only two filters in order to produce, as output, a 2-channel tensor, the two channels encoding for the estimated ToF and stereo confidence respectively. To produce an output with the same resolution of the input, no pooling layers are used. At the same time, to cope with the size reduction at the boundaries due to the convolution operation, each convolutional layer applies a suitable padding to its input along each spatial dimension, where padded values are set to be equal to the values at the boundary.

4.1. Training of the Convolutional Neural Network

The proposed architecture has been trained on the synthetic dataset described in Section 6.1. Although this dataset is smaller if compared with other machine learning datasets, it is the largest dataset for ToF and stereo data fusion containing depth ground truth depth information. We decided to train the network on patches randomly selected from the various scenes instead of using whole images in order to increase the number of training examples. In particular, we generated a large set of training examples by randomly extracting 30 patches of size 128×128 [pxl] from each of the 40 scenes contained in the training set. Moreover, to increase the robustness and variability of the training data, we also augmented the dataset by applying random rotations of $\pm 5^\circ$ as well as horizontal and vertical flipping. Following the augmentation process, a set of about 6000 patches has been generated starting from the 1200 patches initially extracted from the original dataset, thus forming the actual input data used for the training. The training data has been further split into a training set and a validation set. Validation data has been used to select the network layout and parameters. Some ablation studies and results obtained with different network architectures are presented in Section 7.4, in general deeper and more complex architectures led to a smaller training error but there is no improvement in the validation error and in the fusion results due to overfitting on the not too large training dataset.

The two target confidence maps needed for training have been derived by taking the negative exponential of the absolute error gap between the ground truth depth information converted to disparity values D_{GT} and the ToF and stereo disparities D_T and D_S respectively, according to the following formulation:

$$\begin{aligned} P_T^* &= e^{-|D_T - D_{GT}|} \\ P_S^* &= e^{-|D_S - D_{GT}|} \end{aligned} \tag{4}$$

The network has been trained to minimize a canonical quadratic loss function computed as the Mean Squared Errors (MSE) between the predicted ToF and stereo confidence maps P_T and P_S and their corresponding target confidences from Equation (4), i.e.

$$Loss = \sum (P_T - P_T^*)^2 + \sum (P_S - P_S^*)^2 \tag{5}$$

where two summations are taken over the spatial dimensions. Using a single network minimizing a loss function that combines both ToF and stereo error provided better results than training two separate networks to infer ToF and stereo confidences separately.

The optimization has been performed with the AdaDelta algorithm [47]. The process has been carried out using a batch size of 32 and an initial learning rate equal to 0.01. In each convolutional layer, the kernel weights have been initialized following the procedure proposed by He et al. in [46], while all bias values have been initially set to zero.

Both the CNN model as well as the whole optimization and evaluation framework have been implemented using the TensorFlow library [48]. The training stage runs for 500 epochs and takes about 8 hours on a desktop PC with an Intel i7-4790 CPU and an NVIDIA Titan X (Pascal) GPU.

5. Fusion of Stereo and ToF Disparity

The confidence estimated by the deep learning framework of Section 4 can be used to combine the two depth fields coming from the two sensors. The fusion of the upsampled ToF data with the stereo disparity is performed using an extended version of the Locally Consistent (LC) approach.

This method was firstly introduced in [36] for the refinement of stereo disparity data. It refines the disparity estimation by propagating, within an active support centered on the considered point f, the plausibility $\mathcal{P}_{f,g}(d)$ of the disparity assignment coming from other points g inside the active support. The plausibility of a disparity hypothesis d depends on the color and spatial consistency of the considered pixels:

$$\mathcal{P}_{f,g}(d) = e^{-\frac{\Delta_{f,g}}{\gamma_s}} \cdot e^{-\frac{\Delta^{\psi}_{f,g}}{\gamma_c}} \cdot e^{-\frac{\Delta^{\psi}_{f',g'}}{\gamma_c}} \cdot e^{-\frac{\Delta^{\omega}_{g,g'}}{\gamma_t}} \tag{6}$$

where f, g and f', g' refer to the coordinates in the left and right image respectively, Δ accounts for spatial proximity, Δ^{ψ} and Δ^{ω} encode color similarity, and the parameters γ_s, γ_c and γ_t control the relative relevance of the various terms (a detailed description can be found in [36]). The overall plausibility $\Omega_f(d)$ of a disparity hypothesis d is computed by aggregating the plausibility for the same disparity value propagated from neighboring points, i.e.:

$$\Omega_f(d) = \sum_{g \in \mathcal{A}} \mathcal{P}_{f,g}(d). \tag{7}$$

Finally a winner-takes-all strategy is used to compute the optimal disparity value.

A first extension of the approach has been presented in [5] to account for multiple disparity hypotheses as in the case of our setup. The approach of [5] allows to obtain quite good results in the fusion of the two disparity fields but has the key limitation that assigns the same weight to the two data sources without accounting for their reliability.

For this reason the method has been further extended in [6] by assigning different weights to the plausibilities according to the estimated confidence value for each depth acquisition system computed at each pixel location g:

$$\Omega'_f(d) = \sum_{g \in \mathcal{A}} \Big(P_T(g)\mathcal{P}_{f,g,T}(d) \ + \ P_S(g)\mathcal{P}_{f,g,S}(d) \Big) \tag{8}$$

where $\Omega'_f(d)$ is the plausibility at point f for disparity hypothesis d, $\mathcal{P}_{f,g,T}(d)$ is the plausibility propagated by neighboring points g according to ToF data and $\mathcal{P}_{f,g,S}(d)$ is the one according to stereo data. Finally $P_T(g)$ and $P_S(g)$ are the ToF and stereo confidence values at location g respectively. Another improvement to the LC method introduced in [6] is the depth estimation at subpixel precision that allows to obtain a better accuracy. In [6] the confidence information is computed with a deterministic algorithm based on the noise model for the ToF sensor and on the cost function analysis for the stereo system, while in the proposed approach the confidence is estimated with the deep learning architecture of Section 4. For the experimental results of this work the parameters have been set to $\gamma_s = 8$, $\gamma_c = 6$ and $\gamma_t = 4$. Finally notice how the proposed framework can easily be extended to setups with more than two input channels in order to perform the fusion of multiple sensors based on different technologies.

6. Stereo and ToF Datasets

To train the deep network and to evaluate the performance of the proposed approach we acquired two different datasets. The first one, *SYNTH3*, is a synthetic dataset containing 55 different scenes with very different characteristics. The second one, *REAL3*, contains a small number of scenes acquired with a real world trinocular setup made by a ZED stereo camera and a Microsoft Kinect v2 ToF depth camera.

6.1. Synthetic Dataset

The first dataset we built is the *SYNTH3* synthetic dataset [1], which has been generated using the *Blender* 3D rendering software [7]. We downloaded a set of 3D *Blender* scenes from the *Blend Swap* website [49] that have been appropriately modified and rendered from virtual cameras viewpoints in order to generate the stereo-ToF dataset. The virtual acquisition setup is made of a stereo system with characteristics resembling the ones of the ZED stereo camera [50] and a ToF camera with characteristics similar to a Microsoft Kinect v2 [51, 2]. The complete system is depicted in Figure 3, while Table 1 summarizes the parameters of the cameras. More in detail:

Stereo vision system The color images have been generated using *Blender* and the *3D* renderer *LuxRender* [8]. The stereo setup is made of two Full-HD (1920 × 1080) color cameras with a baseline of 12 *cm* and the optical axes and image planes parallel to each other. Notice that these parameters resembles the ones of the ZED camera from Stereolabs used in the real world dataset. Since the cameras are ideal and their optical axes are already aligned there is no need to rectify the two color views, as instead is done for the real world data.

ToF depth camera The data captured by the ToF camera have instead been computed by using the *ToF-Explorer* simulator developed by Sony EuTEC. The *ToF-Explorer* simulator is an extended version of the simulator from Heidelberg University [9] that is able to accurately simulate the data acquired by a real ToF camera including different sources of error as shot noise, thermal noise, read-out noise, lens effect, mixed pixels and the interference due to the global illumination (multi-path effect). The ToF simulator takes in input the scene information generated by *Blender* and *LuxRender*. We acquired with the simulator the 512×424 [pxl] depth and amplitude maps (the resolution and the other simulator parameters have been set in order to emulate the Kinect$^{\text{TM}}$ v2 camera used in the real world setup). The image plane and optical axis of the ToF camera are parallel to those of the stereo camera and the ToF viewpoint is placed under the right camera of the stereo system at a distance of 4 *cm*.

Moreover, the dataset contains also the scene depth ground truth relative to the point of views of the ToF camera and right color camera of the stereo system.

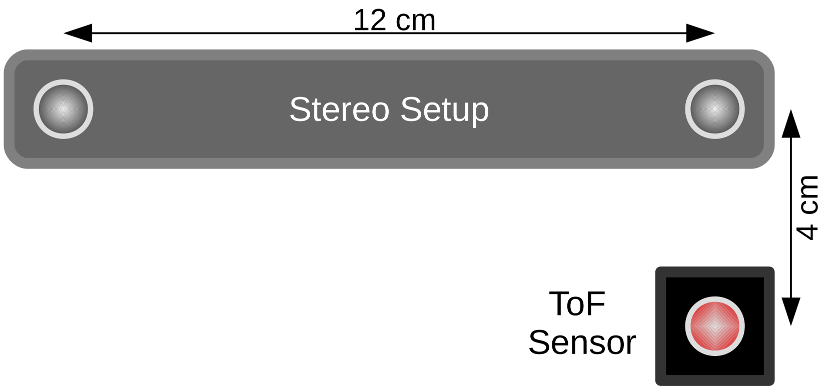


Figure 3: Representation of the synthetic Stereo-ToF acquisition system. The ToF sensor is placed below the color camera.

	Stereo setup	ToF camera
Resolution	1920×1080	512×424
Horizontal FOV	$69°$	$70°$
Focal length	$3.2\ mm$	$3.66\ mm$
Pixel size	$2.2\ \mu m$	$10\ \mu m$

Table 1: Parameters of the stereo and ToF sensors.

The training set contains 20 unique scenes each rendered from 2 different viewpoints, leading to a total of 40 scenes split into a training and a validation set. Even if the number of scenes is low if compared with datasets used for the training of deep networks for other tasks, it is still the largest dataset for stereo-ToF fusion currently available. Furthermore, the scenes are very different one from the other representing different conditions. The test set instead contains 15 unique scenes. The various scenes contain walls, furniture and objects of various shapes and color in different environments, e.g., living rooms, kitchen rooms or offices but also outdoor locations with non-regular structures. The depth range is also very different across the various scenes ranging from about 50 *cm* to 10 *m* thus providing a large range of measurements. The dataset is publicly available at *http://lttm.dei.unipd.it/paper_data/deepfusion*.

6.2. Real World Dataset

The second dataset, *REAL3*, is a real world dataset acquired with a Microsoft Kinect v2 depth camera and a ZED stereo camera. In addition, it also contains the ground truth information generated from the left stereo camera point of view.

We decided to use consumer depth cameras as opposed to expensive professional equipments. In particular the depth cameras used in the collection are:

Stereo vision system We used the ZED camera from Stereolabs [50]. This depth camera based on a passive stereo technology is equipped with two 4MP cameras that provide images up to 2208 × 1242 [pxl] at 15 fps. The sensor is able to provide images up to 100 fps at a lower resolution. The baseline is of 120 [mm] and the diagonal field of view is 110°.

ToF depth camera One of the best consumer ToF depth cameras is the Kinect$^{\text{TM}}$ v2. Compared to other ToF cameras it provides a cleaner and denser depth map and is also the consumer ToF camera with the largest resolution. The Kinect$^{\text{TM}}$ v2 is able to acquire a 512×424 [pxl] depth map at 30 [fps] with a depth estimation error typically smaller than 1% of the measured distances and a diagonal field of view of 92°.

The dataset contains 8 scenes all including static scenarios in an indoor environment. The scenes have different complexity, ranging from flat surfaces to more complex shapes like the leaves of a plant. We acquired objects with and without texture as well in order to check the behavior of the algorithms with disparate conditions. The scenes contain materials with challenging reflection properties, including reflective and glossy surfaces as well as rough material that usually cause problems to active cameras. Figure 4 shows the relative position of the two sensors.

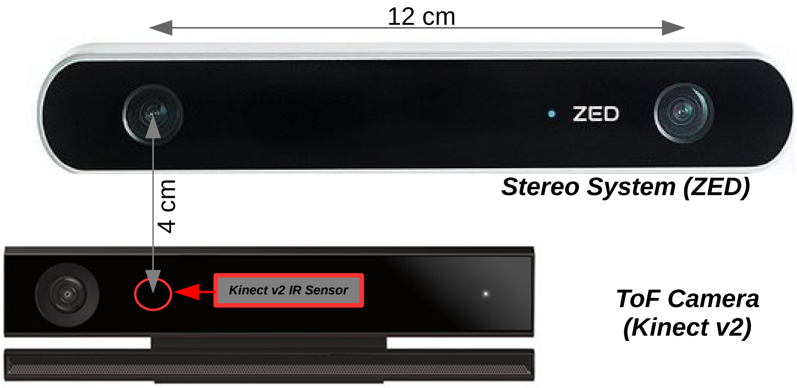


Figure 4: Representation of the real Stereo-ToF acquisition system. The Figure shows the relative position of the ZED camera and of the Kinect v2.

The algorithm developed to compute the ground truth map uses the stereo camera to match corresponding pixels and estimate the disparity between them. We used a line laser with a regular red illuminator visible to humans, acquired by the passive stereo vision system. The goal is to “paint” the scene with the line laser and for each acquisition match corresponding lit points in the two images. Ideally we want to match only 1 point for each row of the image for each acquisition. Due to noise in the images we update the estimated disparity for a given pixel, every time there is a new measurement, by accumulating all the values and keeping the median value. We collected images of the line laser without external illumination to reduce the noise of the acquired images and to increase the contrast of the line laser with respect to the background illumination. To avoid casting unwanted shadows in the scene, the line laser was kept as close as possible to the acquiring cameras. To control the laser movement we used a servomotor that makes the system fully automatic. The dataset is available at *http://lttm.dei.unipd.it/paper_data/realfusion* .

7. Experimental Results

The proposed approach has been evaluated on three different datasets. We started by evaluating its performance on the *SYNTH3* synthetic dataset and then moved to the experiments using data collected with real cameras. For the real world experiments we used both the *REAL3* dataset introduced in this paper and the *LTTM* dataset from [31].

7.1. Evaluation on Synthetic Data

For this set of experiments, the proposed fusion algorithm has been trained and evaluated on synthetic data from the *SYNTH3* dataset described in Section 6.1 (some sample scenes are shown in Figure 5). The *SYNTH3* test set contains 15 different scenes with very different properties including different acquisition ranges, textured and un-textured surfaces, complex geometries and strong reflections. The algorithm takes in input the 512×424 [pxl] depth and amplitude maps from the ToF sensor and the two 960×540 [pxl] color images from the cameras (the color cameras resolution has been halved with respect to the original input data). The output is computed from the point of view of the right camera at the color data resolution of 960×540 [pxl]. For performance evaluation, it has been cropped to consider only on the region that is framed by all the three cameras and compared with ground truth data. Ground truth information has been computed by extracting the depth data from the *Blender* rendering engine and converting it to the disparity space.

Before evaluating the performance of the fusion scheme we analyze the confidence information computed with the deep learning approach of Section 4 that will be used to control the fusion process. Figure 6 shows the color image and the confidence maps for a few sample scenes. The second column shows the ToF confidence, the proposed approach is able to assign a low confidence (darker pixels in the figure) to the areas with a larger error. A first observation is that in most of the confidence maps the error is larger in proximity of the edges. It is a well-known issue of ToF sensors due to the limited resolution and to the mixed pixels effect. Furthermore the CNN is also able to detect that the ToF error is higher on dark surfaces due to the lower reflection (e.g., on the dark furniture in row 3). The multi-path is more challenging to be detected however, by looking at the fruits in row 4, it is possible to see that the confidence is lower in their bottom part touching the

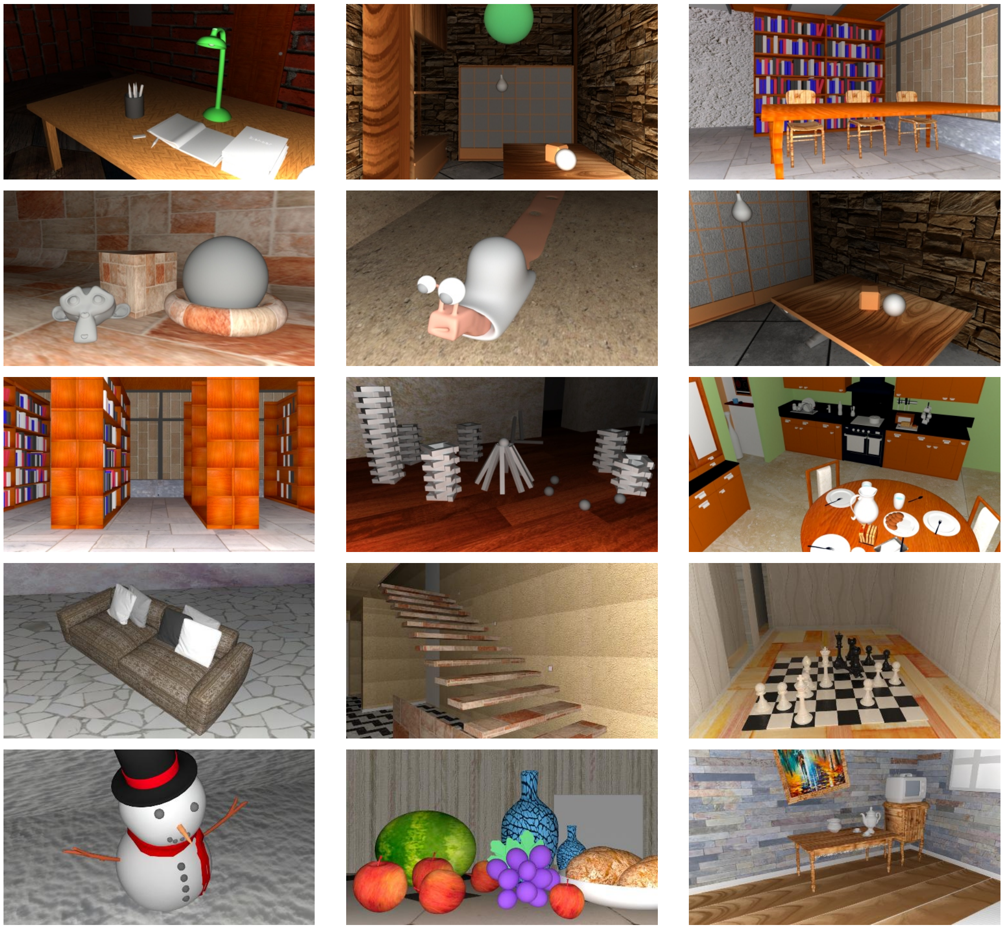
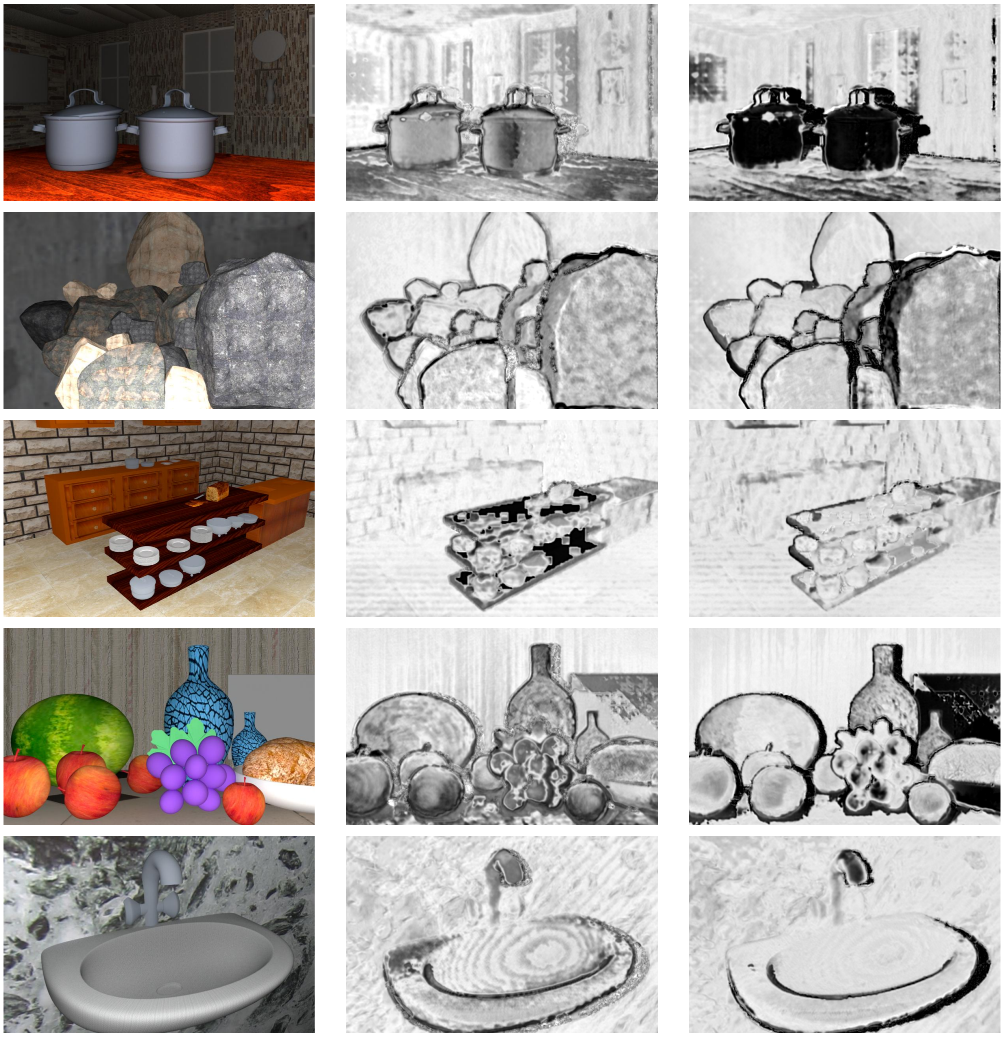


Figure 5: Sample scenes in the *SYNTH3* dataset. The first 3 rows show scenes from the training set while the last 2 from the test set. The figure shows the right camera color image for each scene.



a) Color view b) ToF confidence c) Stereo confidence

Figure 6: Confidence information estimated by the proposed method for some sample scenes: a) color view; b) estimated ToF confidence; c) estimated stereo confidence. Brighter areas correspond to higher confidence values, while darker pixels to lower ones.

table, similarly in the angle between the wall and the sink in row 5, where the multi-path is generated by rays bouncing from one surface to the other.

Concerning the stereo confidence, results are also good. As in the previous case, the limited accuracy on edges is correctly recognized. Furthermore, surfaces with uniform patterns (e.g., the flat panel on the right in row 4) or reflective ones (e.g., the pots in row 1) have lower confidence as expected.

The confidence information is then used to drive the fusion algorithm. The output disparity for some sample scenes is shown in Figure 7. Column 1 shows a color view of the scene while column 2 contains the ground truth disparity data. The up-sampled, filtered and reprojected ToF data are shown in column 3 while column 4 contains the corresponding error map. Notice how ToF data are in general more accurate than the stereo one although some limitations of ToF sensors are visible. In particular, the data in proximity of edges are not too accurate. Furthermore the acquisition on low-reflective surfaces is more noisy and the multi-path error affects some regions close to boundaries between touching surfaces.

Columns 5 and 6 show the disparity and the error map for the SGM stereo vision algorithm. For this work we used the OpenCV implementation of the SGM stereo algorithm with pointwise Birchfield-Tomasi metric, 8 paths for the optimization and a window size of 7×7 [pxl]. Edge regions are challenging also for stereo vision even if they are more accurate than the ToF acquisitions due to the higher resolution. On the other side, some regions proved to be critical for the stereo algorithm, e.g., regions with a limited amount of texture (like the flat panel on the right in row 4) or strongly reflective regions (e.g., the pots in row 1).

Finally, the fused disparity maps and their relative error are shown in columns 7 and 8. The fusion algorithm is able to extract the most accurate information from both sources and provides depth maps with less artifacts on edges but at the same time free from the various artifacts of the stereo acquisition.

The numerical evaluation of the performance is shown in Table 2 and confirms the visual evaluation. The table shows both the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) in disparity space averaged on all the 15 scenes. For a fair comparison, we considered as valid pixels for the results only the ones having a valid disparity value in all the compared disparity maps (stereo, ToF and fused disparities). By looking at the averaged MSE values, the ToF sensor has a high accuracy with a MSE of 4.75, much smaller than the MSE of 13.54 of the stereo system. The MAE

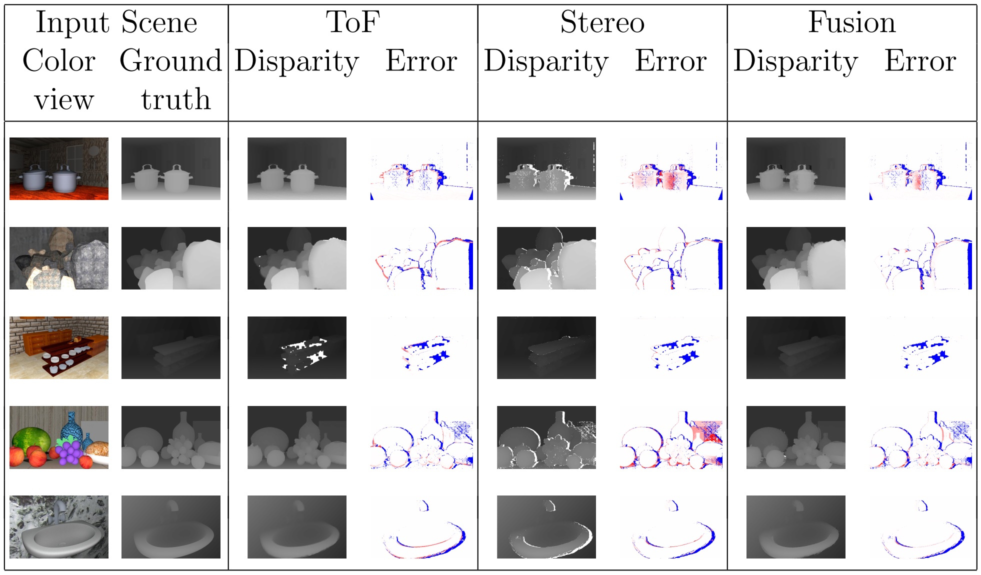


Figure 7: Results of the proposed fusion framework on 5 sample scenes (one for each row). In error images, grey pixels correspond to points excluded since they are not valid on one of the disparity maps. The intensity of red pixels is proportional to the absolute error. (*Best viewed in color*).

Method	*MAE*	*MSE*
Interpolated ToF	0.66	4.75
SGM Stereo	0.79	13.54
Marin et Al. [6]	0.64	4.20
Proposed Method	**0.53**	**3.92**

Table 2: MAE and MSE in disparity units with respect to the ground truth for the ToF and stereo data, the proposed method and [6] on the *SYNTH3* dataset. The error has been computed only on non-occluded pixels for which a disparity value is available in all the methods.

is 0.66 and 0.79 respectively, with a more limited gap due to the fact that the stereo system has some large errors that have a larger impact with the squared measure.

However, confidence data allow to select at most pixel locations the best source and thus to exploit the strengths of both stereo and ToF acquisitions. The proposed approach is able to obtain a MSE of 3.92 and a MAE of 0.53, a very good result with a noticeable improvement with respect to both sensors. Comparison with state-of-the-art approaches on this dataset is limited by the lack of available implementations of the competing approaches. However, we compared our approach with the highly performing method of Marin et Al. [6]. This approach has a MSE of 4.20, higher than the one of the proposed method. The method of [6] outperforms most state-of-the-art approaches, so also the performance of the proposed method are expected to be competitive with the better performing schemes, as demonstrated by the comparison on the *LTTM* dataset in Subsection 7.3.

7.2. Evaluation on Real World Data: REAL3 *dataset*

The testing on synthetic data does not take into account all the potential issues that can arise when working with real world data and sensors. For this reason, we tested the proposed approach also on real world data using the *REAL3* dataset presented in Subsection 6.2. Notice that, due to the limited size of the real world dataset, in this experiment we used the network trained on the synthetic dataset to compute the confidence maps used to drive the fusion process. As pointed out Subsection 6.2, the real world dataset contains 8 different scenes (see Figure 8 for their thumbnails). The scenes are simpler than the synthetic ones due to the challenges in practical data acquisition (specially for what concerns the acquisition of ground truth information), however they contain regions with different amount of texture information, repeating patterns critical for stereo approaches, different materials, bright and dark objects and some complex geometries (e.g., in the plant scene). Similarly to the synthetic data case, the algorithm takes in input the 512×424 [pxl] depth and amplitude maps from the Kinect v2 sensor and the two 960×540 [pxl] color images obtained by subsampling by a factor of 2 and rectifying the two color views from the ZED camera (see sections 3 and 6.2). In this case the output is computed on the point of view of the left camera at the 960×540 [pxl] resolution of color data and compared with the ground truth from the same viewpoint. The estimated disparities have also been cropped to highlight only the region that is framed by all the three cameras.

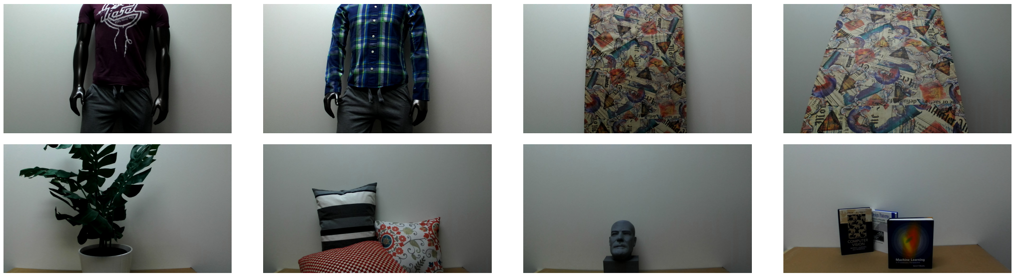
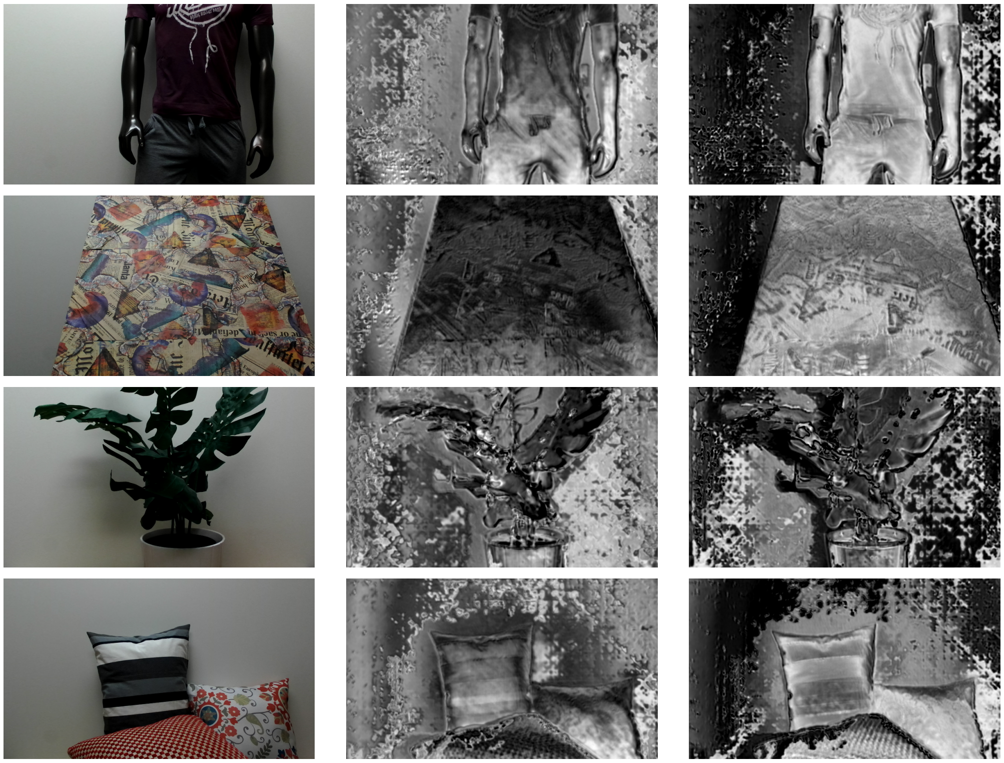


Figure 8: Real world dataset used for the evaluation of the performance of the proposed method. The figure shows the left camera color image for each scene in the dataset.

We start the evaluation from the confidence information: in this case, the task is more challenging since the CNN is trained on synthetic data and then evaluated on the real data, which have slightly different properties. However, the proposed deep network proved to have good generalization properties and the estimated confidence, although not as precise as in the synthetic case, is able to underline the key sources of error as can be seen from the examples in Figure 9. Confidence information for ToF data is shown in the second column, it is possible to note that the CNN properly predicts a higher ToF error in proximity of edges. Also other critical aspects are properly identified, for example in row 2 it is possible to see how black areas in the pattern have a weaker reflection and lead to a less accurate acquisition while in row 3 the CNN properly detects that the acquisition of the plant is very critical for the ToF sensor. The third column contains the confidence maps for stereo data, notice how the CNN is able to recognize that highly textured regions are properly acquired while uniform surfaces like the white walls are critical for the stereo algorithm.

The numerical results of the fusion algorithm are reported in Table 3 while Figure 10 shows the output depth maps and the error maps for some sample scenes. The figure is organized as in the previous experiment. Column 1 and 2 show a color view of the scene and the ground truth disparity data. The up-sampled, filtered and reprojected ToF data are shown in column 3, while column 4 contains the corresponding error map. It is possible to notice that also in this case ToF data are not too precise in proximity of edges, but there is a small amount of error also on flat surfaces due to the noise of the sensor and to inaccuracies in the reprojection operation (in this case it is based on calibration information while previously the cameras were ideally placed).

Columns 5 and 6 show the disparity estimated by the stereo vision algo-



a) Color view b) ToF confidence c) Stereo confidence

Figure 9: Confidence information computed by the proposed deep learning architecture for some sample scenes: a) Color view; b) Estimated ToF confidence; c) Estimated stereo confidence. Brighter areas correspond to higher confidence values, while darker pixels to low confidence ones.

rithm and the corresponding error map. Stereo data have sharper edges and a good accuracy on the objects in the foreground but there are artifacts on low-textured regions, specially on the white walls on the background.

The fused disparity map and its relative error are shown in columns 7 and 8. The fusion algorithm reliably fuses the information coming from the two sensors being able to properly reconstruct the edges using the stereo data but also correctly estimating the background that instead is better acquired by the ToF sensor.

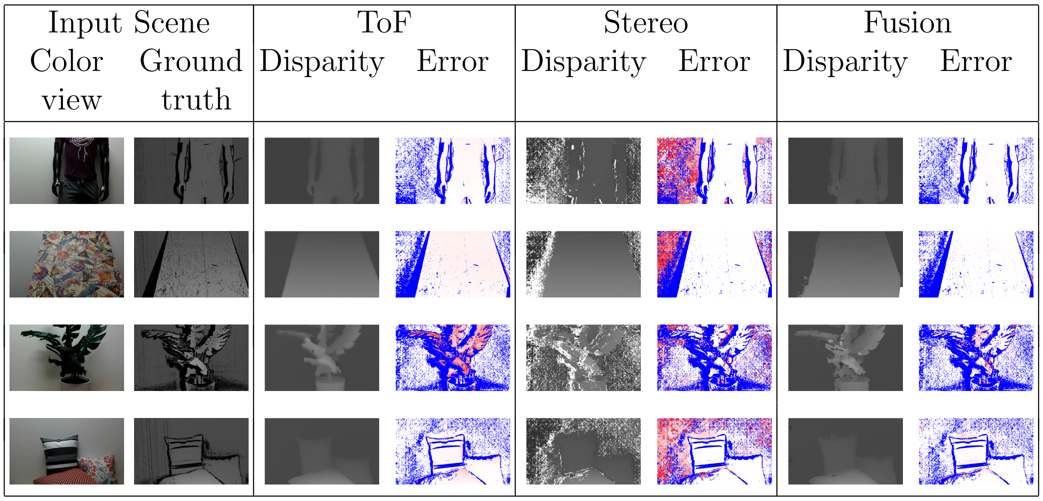


Figure 10: Results of the proposed fusion framework on some sample scenes from the *REAL3* dataset. In the error images, grey pixels correspond to points excluded since they are not valid on one of the disparity maps. The intensity of the red pixels is proportional to the absolute error. (*Best viewed in color*).

The numerical evaluation of the performance is shown in Table 3 and confirms the visual analysis. The table shows the MAE and the MSE in disparity space averaged on all the 8 scenes. For a fair comparison, we considered as valid pixels for the results only the ones having a valid disparity value in all the compared disparity maps (stereo, ToF and fused disparities). By looking at the MAE values, the ToF sensor has a high accuracy with an error of 2.55, much smaller than the MAE of 7.98 of the stereo system (and the MSE difference is even larger). This is a challenging situation for fusion algorithms since it is difficult to improve the data from the best sensor without affecting it with errors from the other one. However, the use of confidence data helps to properly combine both sources of information obtaining a MAE of 1.65 with a noticeable improvement with respect to both sensors. The method of Marin et Al. [6] on this dataset has a MAE of 2.19, again higher than the one obtained with the proposed method.

Method	*MAE*	*MSE*
Interpolated ToF	2.55	10.76
SGM Stereo	7.98	201.64
Marin et Al. [6]	2.19	8.82
Proposed Fusion	**1.65**	**8.35**

Table 3: MAE and MSE in disparity units with respect to the ground truth for the ToF and stereo data, the proposed method and [6] on the *REAL3* dataset. The error has been computed only on non-occluded pixels for which a disparity value is available in all the methods.

7.3. Evaluation on Real World Data: LTTM *dataset*

Finally, we tested the proposed approach on the *LTTM* dataset. This dataset has been introduced in [31] and contains 5 different scenes acquired with a MESA SR4000 ToF sensor and two Basler color cameras (the scene thumbnails are in the first column of Figure 11). Even if it is smaller than the other two datasets and the ToF data has been acquired with a camera with lower performance (the resolution is just 176 × 144 [pxl]), this dataset represents an interesting benchmark since it has been used for the evaluation of several works and allows to to perform the comparison with different state-of-the-art methods from the literature. Furthermore it contains object with various shapes and characteristics that allow to evaluate the method in various situations including depth discontinuities, materials with different reflectivity and both textured and un-textured surfaces. In order to process this dataset the algorithm takes in input the 176 × 144 [pxl] depth and amplitude maps from the MESA sensor and the two 1032 × 778 [pxl] color images from the Basler cameras and computes the output from the point of view of the left camera at the same resolution of color data. For confidence estimation we used the CNN trained on synthetic data from the *SYNTH3* training set as for the other datasets.

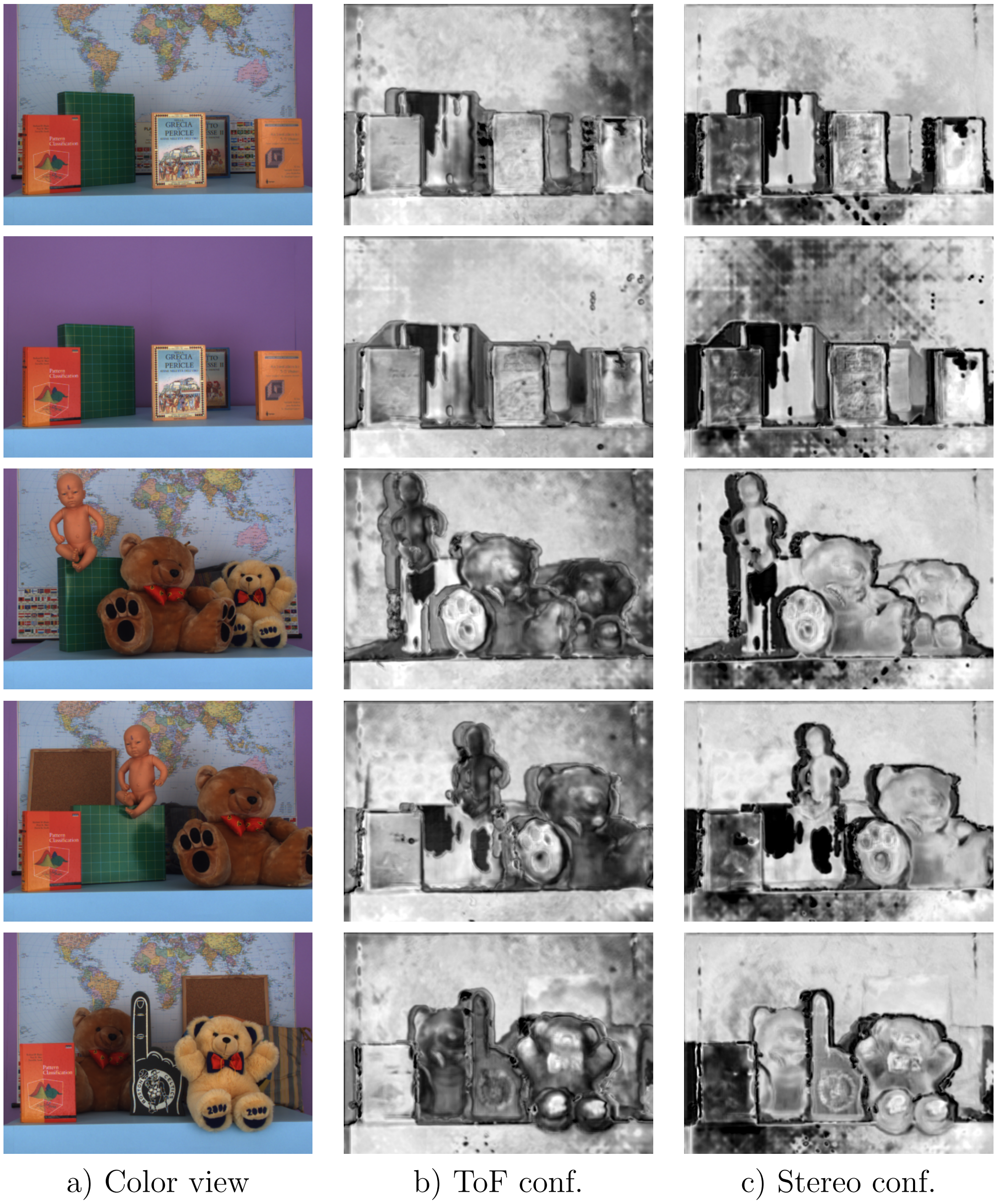
Figure 11 shows the confidence information for the ToF and stereo sensors on this dataset. This situation is even more challenging since the ToF camera used for this dataset has very different properties from the simulated one used in the training. The accuracy of confidence information is lower, however the proposed approach is able to detect some key issues. Concerning ToF data it is possible to notice the lower confidence in proximity of edges and that the depth information is less reliable on the complex geometries of the objects if compared with the walls and table. Stereo data are also less reliable on edges and on regions with a lower amount of texture.

Concerning the results of the fusion of the two sensors, Figure 12 shows the output depth maps and the error maps for the 5 scenes of the dataset. It is possible to notice the good accuracy of fused data on edges and how the algorithm is able to properly choose the best data source in many situations avoiding the artifacts of the two acquisition devices. For example, the repeating pattern on the green box causes errors in the stereo reconstruction that are not present in the fused data. On the other side, the upper part of the table is very critical for the ToF sensor due to the multi-path and to the surface orientation. In the fused disparity, even if not perfect, it is better reconstructed thanks to the information coming from stereo vision.

Table 4 reports the numerical values for the error and the comparison with some state-of-the-art methods from the literature.

The compared state-of-the-art methods are based on different strategies: the method of [38] uses an iterative approach and bilateral filtering. Then we considered two approaches based on probabilistic MAP-MRF schemes, i.e., [27] and [31]. Finally, there are the two previous approaches based on the LC framework, i.e. [5] and [6]. For a fair comparison, we considered as valid pixels for the results only the ones having a valid disparity value in all the compared disparity maps (stereo, ToF and fused disparities from the various methods). By looking at the average values, on this dataset the ToF and stereo sensors have a similar MAE of 1.53 and 1.45 while the MSE is much lower for the ToF sensor (this is due to the fact that the stereo data has some large errors while ToF error is more uniformly distributed).

The proposed approach achieves a MAE of 0.89, that is better than all the proposed approaches with a large margin. The best among the compared approaches is [6], that has a MAE about 25% higher, while all the other compared approaches have a larger error. If using the MSE as error metric the gap with [6] is smaller while it remains large with respect to all the other approaches. This is due to the fact that [6] relies strongly on ToF data that has a better MSE while the proposed approach makes a more balanced use of the two sources of information. In any case, the proposed approach has the best performance among all the compared ones according to both measures.



a) Color view

b) ToF conf.

c) Stereo conf.

Figure 11: Confidence information computed by the proposed deep learning architecture for the scenes in the *LTTM* dataset: a) Color view; b) Estimated ToF confidence; c) Estimated stereo confidence. Brighter areas correspond to higher confidence values, while darker pixels to low confidence ones.

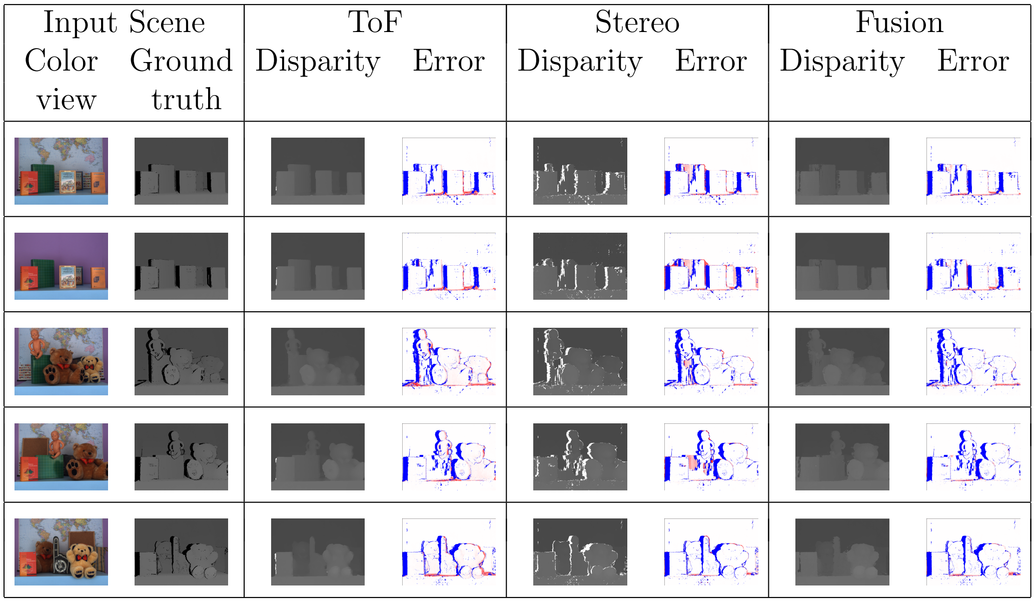


Figure 12: Results of the proposed fusion framework on the 5 scenes of the *LTTM* dataset. In error images, grey pixels correspond to points excluded since they are not valid on one of the disparity maps. The intensity of red pixels is proportional to the absolute error. (*Best viewed in color*).

Method	*MAE*	*MSE*
Interpolated ToF	1.53	11.68
SGM Stereo	1.45	20.42
Dal Mutto et Al. (LC) [5]	1.36	10.06
Marin et Al. [6]	1.15	7.67
Yang et Al. [38]	1.59	10.98
Zhu et Al. [27]	1.59	11.13
Dal Mutto et Al. (MRF) [31]	1.43	12.21
Proposed Fusion	**0.89**	**7.40**

Table 4: MAE and MSE averaged on the 5 scenes of the *LTTM* dataset (in disparity units) with respect to the ground truth, computed only on non-occluded pixels for which a disparity value is available in all the methods.

7.4. Ablation Studies

Finally, we performed some further tests in order to evaluate the impact on the fusion accuracy of the information coming from the various input sources. Specifically, we made an additional set of experiments where, in turn, we selectively removed one of the four input sources of Section 4 in order to better evaluate its contribution to the final output. The results are shown in Table 5 and indicate how, on average, the combination of all inputs offers the best performance. More in detail, the first two rows show how each of the two disparities contains relevant information for the corresponding sensor. Both the removal of the ToF or stereo disparity leads to a quite large decrease in terms of fusion accuracy (around 20%). The impact of the ToF amplitude is smaller, but it has a noticeable effect on the *SYNTH3* and *LTTM* datasets. The main issue with ToF amplitude is that it depends a lot on the employed sensor while the ToF simulator is not able to model in a completely accurate way the amplitude data acquired by real world sensors. Finally, the difference map Δ between the reference image and the target one reprojected over it proved to be very useful specially in real world datasets where the stereo matching is less reliable. Concluding, even if not all information types are fundamental for all datasets, the combination of all the four sources is the best solution in order to have an approach with very good performance on both real and synthetic data.

Ablation Study	*SYNTH3*	*REAL3*	*LTTM*
All inputs (proposed method)	0.53	1.65	0.89
Without ToF disparity (no D_T)	0.64	2.04	1.179
Without Stereo disparity (no D_S)	0.66	2.11	1.19
Without ToF amplitude (no A_T)	0.55	1.6	0.915
Without LR difference (no Δ)	0.52	2.84	1.19
Separate estimation ToF/stereo conf.	0.61	1.93	1.156
Select highest confidence (HC)	0.43	1.90	1.11
Weighted average (WA)	0.46	2.44	1.07

Table 5: Mean Absolute Error (MAE) on the fused depth maps (in disparity units) when removing different input channels, with separate stereo and ToF confidence estimation and with different fusion strategies.

Another idea we exploited in the paper is the one of jointly estimating the confidence of the two sensors instead of independently computing the two confidence maps. We evaluated the impact of this approach by trying to estimate the ToF and stereo confidence separately with two different CNNs with a single output (keeping fixed the other parameters). As shown by the last row of the table, this approach leads to worse performances on all 3 datasets, demonstrating that the joint estimation allows to obtain more coherent confidence maps and thus a better accuracy of the fused data.

Finally, in order to evaluate the impact of the Locally Consistent (LC) fusion algorithm we tried also to exploit the confidence data estimated with the proposed deep learning architecture into simpler fusion strategies. We tried two simple solutions, the first is the selection at each location of the source with the highest confidence (HC) and the second is the usage of a weighted average (WA) of the ToF and stereo disparities with the weights given by the estimated confidences at each pixel location. The obtained results are in the last two rows of Table 5. On synthetic data, where the confidence information is very reliable and the noise on the data is limited, even by just selecting at each pixel location the source with the highest estimated confidence it is possible to obtain very good results with a MAE of 0.43, even better than the one achieved by the LC algorithm. Also the weighted average driven by confidence allows to obtain a very good result with a MAE of 0.46. This proves the reliability of the proposed confidence estimation algorithm. While on synthetic data the LC refinement is not really necessary, the discussion is quite different on real world data. On the *REAL3* dataset the selection of the source with the highest confidence and the weighted average achieve a MAE of 1.9 and 2.44 respectively, quite higher than the result of the full version of the proposed approach with LC (which achieves a MAE of 1.6). A similar discussion holds for the *LTTM* dataset (the absolute errors are 1.11 for HS and 1.07 for WA against 0.89 for LC). This proves how the smoothing and regularization of the choices performed by LC is very useful when data are more noisy and less reliable as it happens in real world acquisitions. On the other side, simpler fusion strategies might be preferable when a fast computation is needed or data are very reliable.

8. Conclusions and Future Work

In this work we presented a scheme for the fusion of ToF and stereo data using confidence information estimated by a deep learning architecture. We created a novel synthetic dataset containing a realistic representation of the data acquired by a passive stereo camera and a ToF depth camera and a second dataset containing data from real depth cameras with the associated ground truth information. Using these datasets we have showed how the confidence estimation network is able to generalize to different datasets with both synthetic and real data.

More in detail, a Convolutional Neural Network trained on the synthetic dataset is used to estimate the reliability of ToF and stereo data, obtaining confidence maps that highlight the most critical acquisition issues of both sub-systems. The fusion of the two sources of depth data is then performed using an extended version of the LC framework that combines the confidence information computed in the previous step and provides an accurate disparity estimation. The results show how the proposed algorithm properly combines the outputs of the two sensors providing, on average, a disparity map with higher accuracy with respect to each of the two sub-systems. The test was performed on 3 different datasets, with both synthetic and real data obtaining very good performances and outperforming state-of-the-art approaches. We believe that showing how it is possible to train a network on synthetic data and obtain comparable performance on real data is of fundamental importance in a data fusion framework.

Further research will be devoted to the improvement of the deep learning architecture with the target of obtaining a more reliable confidence information. In particular, we will consider multi-branch architectures for a better combination of the two data sources. We also plan to develop an end-to-end deep learning architecture to directly compute the final output. Finally, we will also acquire larger datasets for a better training of the CNN and we are going to investigate the fusion of other active depth sensors.

Acknowledgment. We would like to thank Prof. Stefano Mattoccia for providing the source code of the LC framework and for the very valuable help for using it in the sensor fusion application. We also thank Matteo Poggi for the useful discussion and support on the confidence estimation task and the Computational Imaging Group at the Sony European Technology Center (EuTEC) for allowing us to use their *ToF-Explorer* simulator. We also thank Prof. Giancarlo Calvagno for his support and gratefully acknowledge NVIDIA Corporation for the donation of the GPUs used for this research.

References

[1] G. Agresti, L. Minto, G. Marin, P. Zanuttigh, Deep learning for confidence information in stereo and tof data fusion, in: ICCV Workshop: 3D Reconstruction meets Semantics, 2017.

[2] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, G. M. Cortelazzo, Time-of-Flight and Structured Light Depth Cameras: Technology and Applications, 1st Edition, Springer International Publishing, 2016.

[3] M. Poggi, F. Tosi, S. Mattoccia, Quantitative evaluation of confidence measures in a machine learning world, in: International Conference on Computer Vision (ICCV 2017), 2017.

[4] X. Hu, P. Mordohai, A quantitative evaluation of confidence measures for stereo vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2121–2133.

[5] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, G. Cortelazzo, Locally consistent tof and stereo data fusion, in: Workshop on Consumer Depth Cameras for Computer Vision (ECCV Workshop), Springer, 2012, pp. 598–607.

[6] G. Marin, P. Zanuttigh, S. Mattoccia, Reliable fusion of tof and stereo depth driven by confidence measures, in: European Conference on Computer Vision, Springer International Publishing, 2016, pp. 386–401.

[7] Blender website, https://www.blender.org/ (Accessed July 31st, 2017).

[8] Luxrender website, http://www.luxrender.net (Accessed July 31st, 2017).

[9] S. Meister, R. Nair, D. Kondermann, Simulation of Time-of-Flight Sensors using Global Illumination, in: M. Bronstein, J. Favre, K. Hormann (Eds.), Vision, Modeling and Visualization, The Eurographics Association, 2013.

[10] Middlebury stereo vision benchmark, http://vision.middlebury.edu/stereo/ (Accessed October 16th, 2017).

[11] The kitti vision benchmark suite, http://www.cvlibs.net/datasets/kitti/ (Accessed October 16th, 2017).

[12] B. Tippetts, D. Lee, K. Lillywhite, J. Archibald, Review of stereo vision algorithms and their suitability for resource-limited systems, Journal of Real-Time Image Processing (2013) 1–21.

[13] M. Poggi, S. Mattoccia, Learning from scratch a confidence measure, in: Proceedings of the British Machine Vision Conference (BMVC), 2016.

[14] A. Seki, M. Pollefeys, Patch based confidence prediction for dense disparity map, in: Proceedings of the British Machine Vision Conference (BMVC), 2016.

[15] M. Poggi, S. Mattoccia, Learning to predict stereo reliability enforcing local consistency of confidence maps, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[16] M. Hansard, S. Lee, O. Choi, R. Horaud, Time-of-Flight Cameras: Principles, Methods and Applications, SpringerBriefs in Computer Science, Springer, 2013.

[17] F. Remondino, D. Stoppa (Eds.), TOF Range-Imaging Cameras, Springer, 2013.

[18] D. Piatti, F. Rinaudo, Sr-4000 and camcube3.0 time of flight (tof) cameras: Tests and comparison, Remote Sensing 4 (4) (2012) 1069–1089.

[19] T. Kahlmann, H. Ingensand, Calibration and development for increased accuracy of 3d range imaging cameras, Journal of Applied Geodesy 2 (2008) 1–11.

[20] S. A. Gudmundsson, H. Aanaes, R. Larsen, Fusion of stereo vision and time of flight imaging for improved 3d estimation, Int. J. Intell. Syst. Technol. Appl. 5 (2008) 425–433.

[21] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2012, pp. 746–760.

[22] S. Song, S. P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite., in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 5, 2015, p. 6.

[23] A. Memo, P. Zanuttigh, Head-mounted gesture controlled interface for human-computer interaction, Multimedia Tools and Applications 77 (1) (2018) 27–53.

[24] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, R. Singh, Unconstrained kinect video face database, Information Fusion 44 (2018) 113 – 125.

[25] J. Marco, Q. Hernandez, A. Muñoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, D. Gutierrez, Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging, ACM Trans. Graph. 36 (6) (2017) 219:1–219:12.

[26] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, M. Eisemann, M. Magnor, D. Kondermann, A survey on time-of-flight stereo fusion, in: M. Grzegorzek, C. Theobalt, R. Koch, A. Kolb (Eds.), Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Vol. 8200 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 105–127.

[27] J. Zhu, L. Wang, R. Yang, J. Davis, Fusion of time-of-flight depth and stereo for high accuracy depth maps, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[28] J. Zhu, L. Wang, J. Gao, R. Yang, Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 899–909.

[29] J. Zhu, L. Wang, R. Yang, J. E. Davis, Z. Pan, Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (7) (2011) 1400–1414.

[30] C. Dal Mutto, P. Zanuttigh, G. Cortelazzo, A probabilistic approach to ToF and stereo data fusion, in: Proc. of 3DPVT, Paris, France, 2010.

[31] C. Dal Mutto, P. Zanuttigh, G. Cortelazzo, Probabilistic tof and stereo data fusion based on mixed pixels measurement models, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (11) (2015) 2260–2272.

[32] R. Nair, F. Lenzen, S. Meister, H. Schaefer, C. Garbe, D. Kondermann, High accuracy tof and stereo sensor fusion at interactive rates, in: Proceedings of European Conference on Computer Vision Workshops (ECCVW), 2012.

[33] B. Chen, C. Jung, Z. Zhang, Variational fusion of time-of-flight and stereo data using edge selective joint filtering, in: Proceedings of International Conference on Image Processing, 2017.

[34] B. Chen, C. Jung, Z. Zhang, Variational fusion of time-of-flight and stereo data for depth estimation using edge selective joint filtering, IEEE Transactions on Multimedia (2018) 1–1.

[35] G. Evangelidis, M. Hansard, R. Horaud, Fusion of Range and Stereo Data for High-Resolution Scene-Modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (11) (2015) 2178 – 2192.

[36] S. Mattoccia, A locally global approach to stereo correspondence, in: Proc. of 3D Digital Imaging and Modeling (3DIM), 2009.

[37] J. Diebel, S. Thrun, An application of markov random fields to range sensing, in: In Proc. of NIPS, MIT Press, 2005, pp. 291–298.

[38] Q. Yang, R. Yang, J. Davis, D. Nister, Spatial-depth super resolution for range images, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[39] Q. Yang, N. Ahuja, R. Yang, K. Tan, J. Davis, B. Culbertson, J. Apostolopoulos, G. Wang, Fusion of median and bilateral filtering for range image upsampling, IEEE Transactions on Image Processing.

[40] S. Schwarz, M. Sjostrom, R. Olsson, Time-of-flight sensor fusion with depth measurement reliability weighting, in: Proceedings of the 3DTV Conference, 2014, pp. 1–4.

[41] V. Garro, C. Dal Mutto, P. Zanuttigh, G. M. Cortelazzo, A novel interpolation scheme for range data with side information, in: Proceedings of the European Conference on Visual Media Production (CVMP), 2009.

[42] J. Dolson, J. Baek, C. Plagemann, S. Thrun, Upsampling range data in dynamic environments, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 1141–1148.

[43] Z. Zhang, A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1998) 1330–1334.

[44] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603 –619.

[45] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[46] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[47] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701.

[48] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).

[49] Blend swap website, https://www.blendswap.com/ (Accessed July 31st, 2017).

[50] Zed stereo camera, https://www.stereolabs.com/ (Accessed July 31st, 2017).

[51] J. Sell, P. O'Connor, The xbox one system on a chip and kinect sensor, IEEE Micro 34 (2) (2014) 44–53.