Confidence Estimation for ToF and Stereo Sensors and Its Application to Depth Data Fusion

# Confidence Estimation for ToF and Stereo Sensors and its Application to Depth Data Fusion

Matteo Poggi, *Member, IEEE,* Gianluca Agresti, *Student Member, IEEE,* Fabio Tosi, *Student Member, IEEE,* Pietro Zanuttigh, *Member, IEEE,* and Stefano Mattoccia, *Member, IEEE*

***Abstract*—Time-of-Flight (ToF) sensors and stereo vision systems are two widely used technologies for depth estimation. Due to their rather complementary strengths and limitations, the two sensors are often combined to infer more accurate depth maps. A key research issue in this field is how to estimate the reliability of the sensed depth data. While this problem has been widely studied for stereo systems, it has been seldom considered for ToF sensors. Therefore, starting from the work done for stereo data, in this paper, we firstly introduce novel confidence estimation techniques for ToF data. Moreover, we also show how by using learning-based confidence metrics jointly trained on the two sensors yields better performance. Finally, deploying different fusion frameworks, we show how confidence estimation can be exploited in order to guide the fusion of depth data from the two sensors. Experimental results show how accurate confidence cues allow outperforming state-of-the-art data fusion schemes even with the simplest fusion strategies known in the literature.**

***Index Terms*—Time-of-Flight, Stereo Vision, Confidence Information, Data Fusion, Deep Learning.**

## I. Introduction

Depth sensing is a challenging task for which many different devices and approaches exist. However, none of them is completely satisfactory, specially when dealing with dynamic scenes. Structured light scanners are suitable only for the acquisition of static objects in indoor environment while laser-based methods such as LiDAR are generally expensive, cumbersome, based on moving mechanical parts and only partially suited to dynamic environments. Among the various possible solutions for depth estimation, two techniques are increasing their popularity due to their simplicity, low cost and capability of handling dynamic scenes: stereo vision systems and matricial Time-of-Flight (ToF) sensors. Even if widely deployed in several practical applications, both approaches have their specific shortcomings. Stereo vision uniquely rely on images framed by standard imaging devices and thus provides unreliable depth measurements when the matching of corresponding points is ambiguous, like for instance when dealing with low-textured regions. On the other hand, ToF sensors have a limited resolution and suffer from mixed pixels and Multi-Path Interference (MPI) artifacts [1], [2], [3].

Regardless of the depth sensing technique deployed, for the reasons outlined, extracting reliable confidence metrics for depth data is a relevant task. For stereo vision, confidence estimation has been a widely explored research field and recently traditional techniques have been outperformed by a large margin by machine learning approaches. On the other hand, inferring confidence estimation from ToF sensors is an almost unexplored field. Thus, in this paper we will present different methods for this purpose, ranging from simple traditional strategies to the adaptation of machine learning (ML) approaches developed for stereo vision and *ad-hoc* ML strategies explicitly targeted to ToF sensors.

Among the various applications, confidence data proved to be very useful when depth data coming from the two approaches needs to be combined together. For this task it is important to ensure that the two confidence metrics are consistent and we will show that by using deep networks able to jointly estimate both confidence metrics yields the best results. Finally, we will exploit the proposed confidence estimation strategies into different depth data fusion frameworks. An extensive experimental evaluation on three different datasets, including both synthetic and real world scenes, has been carried out for the confidence measures and for the data fusion algorithms. The results show how the proposed deep learning approaches, specially if jointly trained on both devices, allow to obtain a very accurate confidence information. Furthermore, state-of-the-art results on stereo-ToF fusion can be obtained by exploiting the computed confidence data to guide the fusion algorithm.

The rest of the manuscript is organized as follows: Section II will discuss the related literature, Sections III, IV and V will introduce in detail stereo, ToF and joint confidence estimation respectively, Section VI will describe in detail fusion strategies, Section VII will show the experimental results for both standalone confidence estimation and depth fusion, finally Section VIII will draw the conclusions.

We list here a few symbols recurring in the paper. We refer to as $d$ and $z$ for disparity and depth values sourced by stereo and ToF. We denote with $\hat{C}_S$ and $\hat{C}_T$ the estimated confidence maps for stereo and ToF respectively, and as $C_S$ and $C_T$ the corresponding ground truth confidence labels that range from 1 (when the error estimated on $d$ or $z$ is 0) to 0 (when the estimated $d$ or $z$ is not reliable).

## II. Related Works

Depth estimation using stereo vision is a long-standing research field, and a large number of approaches have been proposed and tested on public datasets like the Middlebury [4] and KITTI [5] benchmarks. A comprehensive review of stereo methods can be found in [6]. Despite the research efforts and the continuous improvements, the depth estimation

M. Poggi, F. Tosi and S. Mattoccia are with the University of Bologna.
G. Agresti and P. Zanuttigh are with the University of Padova.

accuracy of a stereo system is primarily affected by specific scene contents. In particular, the estimation is less accurate in poorly textured regions or when the scene contains repetitive patterns. Since the accuracy can vary considerably between different scenes or even different regions of the same scene, it is crucial to estimate the confidence of the inferred depth data. The confidence information for stereo systems has been traditionally estimated by analyzing some key properties of the stereo matching cost function (see [7] for a comprehensive review of approaches based on this strategy).

ToF cameras represent a quite robust solution for depth acquisition and there are many consumer ToF depth cameras able to acquire reliable depth data at interactive frame rates. The technology behind these sensors has been analyzed in various works [8], [9], [3], [10], [11], [12]: they are more independent from the scene content w.r.t. stereo systems since they are active devices and do not rely on the photometric content of the scene but they have other limitations. First of all, the resolution is typically low and the noise level is quite high. The low resolution also causes pixels close to edges to receive light from different regions, thus producing wrong measurements (*mixed pixel* effect). Furthermore, light rays can bounce multiple times in the scene causing an over-estimation of the depth (*Multi-Path Interference*).

The estimation of confidence information for ToF sensors is still a quite unexplored field. Some early works used analytical approaches to estimate the reliability of ToF data [13], [14], [15], [16], [17] typically exploiting ToF noise models (specially the idea that the noise is proportional to the amplitude of the received signal) and the assumption that data is less reliable close to edges.

More recently, machine learning techniques emerged for stereo and ToF confidence estimation. For stereo, confidence has been estimated firstly with classifiers based on Random Forests, then by using deep learning techniques (see [18] for a recent review of learning-based confidence estimation methods). In [19], a Convolutional Neural Network (CNN) estimated confidence from image patches while a two-channel image patch representation is used by [20] for the same purpose. In [21], a deep network improved confidence measures by enforcing their local consistency. Machine learning strategies have also been used for ToF sensors, even if the limited availability of ToF data with depth ground truth has limited their diffusion in this field. An early attempt for ToF confidence estimation exploiting a Random Forest classifier is [22]. Later on, [23] exploited a CNN trained on synthetic data. A related task is ToF data denoising where the reliability estimation is implicitly performed to decide where to apply the refinement strategies. For this task, approaches based on deep learning are [24], [25] and [26].

Since ToF cameras and stereo systems have different and rather complementary shortcomings, the fusion of depth fields coming from the two devices can lead to a more accurate depth estimation and has been the subject of many research works. Recent reviews on this topic can be found in [27] and [3]. A first family of works exploits probabilistic approaches, from simple ML estimators [15] to more advanced schemes based on MAP-MRF Bayesian frameworks [28], [29], [30], [16]. A second family of approaches is based on variational fusion frameworks, e.g., the approach of [17] (that also exploits confidence measures to control the fusion), or the more recent works of [31], [32]. Other solutions exploit a local energy minimization formulation [33], the locally consistent framework [34], [14], or bilateral filtering driven by confidence information [35]. Deep learning for this task has been introduced in [23] and [36], extending the work of [14] by estimating the confidence with a CNN. Finally, the fusion of multiple depth maps with deep learning has been considered [37] in the context of multi-view 3D reconstruction.

## III. Confidence estimation for stereo vision

Confidence estimation [7], [18] has a long history in the field of stereo vision. Traditionally, confidence scores were obtained by studying peculiar cues available from the cost volume processed by stereo algorithms [7]. A popular example is the Peak Ration measure (PKR), consisting into the ratio between the minimum cost and the one closest to it: the larger is such feature, the less ambiguous is the minimum selection for the pixel and thus the more reliable is assumed the disparity assignment. In literature, we find measures leveraging matching costs, local or global properties of the per-pixel cost curve, left-right consistency and distinctiveness [7].

Since different cues better encode different behaviors, each measure has its strengths and weaknesses, resulting particularly effective at detecting unreliable disparities in certain conditions while failing in others. For instance, the left-right consistency check (LRC) is particularly useful in the presence of occlusions but may fail at detecting mismatches due to other challenging situations such as ambiguous or repetitive patterns. Given such orthogonality between different measures, an effective strategy to infer confidence estimation consists in combining multiple measures using machine learning, and in particular random forest classifiers [38], [39], [40], [41]. These works proved that training a machine learning model fed with multiple measures can significantly improve effectiveness at detecting outliers.

Concerning learning-based measures, a particularly appealing strategy, referred to as O1, uses as single input cue only the disparity map [41]. Such an approach is potentially suited with any depth sensor providing dense disparity estimation even when a cost volume is not available (e.g., consumer devices such as the Intel RealSense stereo camera). The O1 approach [41] combines random forest classifiers with disparity-based measures (disparity agreement DA, disparity scattering DS, disparity variance VAR, median disparity MED and median disparity deviation MDD) and features on increasing neighborhood windows proving to be more effective with respect to features set extracted from the cost volume [38], [39], [40].

According to [18], some of the features computed by [41] in the disparity domain, in particular DA and DS, turn out to be particularly effective compared to traditional confidence measures computed from the cost volume. Specifically, for each pixel $p$ and its neighborhood $N$, DA and DS are respectively obtained as the number of pixels sharing the same disparity hypothesis of $p$ and the total number of disparity hypothesis

occurring in $N$. Thanks to their independence from the cost volume, these features can be possibly deployed also to depth maps obtained from different techniques, as in the case of ToF, whereas measures leveraging the cost volume could not.

Finally, in [14] a confidence measure based on the relationship (within the Semi-Global Matching (SGM) framework [42]) between matching costs after local aggregation and global optimization has been proposed (we will denote it as ST-D$_S$). Despite not very effective to detect outliers [18], such measure is particularly good for stereo and ToF fusion [14].

A further step towards better confidence estimation for stereo vision relies on CNN based measures aimed at inferring confidence scores from direct processing of the disparity map as O1. Some CNN-based methods extracting simple features [20] while others *from scratch* [19], [43]. Regardless of the adopted strategy, as for O1 and its features, these techniques are independent of intermediate representations such as matching costs. Therefore, they can process any depth map obtained by other means than stereo.

We can classify deep learning based measures according to the portion of data considered for confidence estimation.

- **Local approaches.** Approaches belonging to this category estimate confidence scores by processing small patches from the disparity map. CCNN [19] and PBCP [20] follow this strategy, respectively processing $9 \times 9$ and $15 \times 15$ windows. While CCNN is computed from a single disparity map alone, PBCP needs both left and right disparity maps to learn a left-right consistency, thus being strongly customed to stereo and not suited for ToF data. Specifically, CCNN consists of four $3 \times 3$ convolutional layers extracting 64 features and further three $1 \times 1$ layers extracting respectively 100, 100 and 1 features. The framework by Agresti et al. [23] belongs to this category as well.
- **Local-Global approaches.** This category combines local information processed by patch-based CNNs with global context extracted by the larger receptive fields of encoder-decoder architectures. LGC [43] effectively carries out this combining CCNN [19] and ConfNet [43]. More specifically, ConfNet is made of $3 \times 3$ layers extracting $64, 128, 256, 512$ features respectively and $2 \times 2$ max-pooling operations adopted in order to decimate deep features. Then $3 \times 3$ deconvolutional layers with stride 2 extract $256, 128, 64, 32$ features and predict a final confidence map restoring the original input resolution. Since the right disparity map is not required, LGC is suited for general purpose depth maps, thus for ToF data as well. In particular, LGC combines in a cascade manner the output of both local and global networks, the depth map and the reference image employing a final module of the same structure of CCNN [19] to improve the final confidence prediction.

## IV. Confidence estimation for ToF sensors

The acquisition of Time-of-Flight depth data is affected by multiple error sources [3], including thermal noise, multi-path interference and the mixed-pixel effect. Confidence estimation should take into account all these issues and how they are related to the scene characteristics. For this task, different strategies can be considered, from simple traditional approaches to learning-based solutions.

### A. Confidence from amplitude and intensity values

The noise on the depth map is strictly related to the amplitude $A$ of the received signal, that depends on the distance of the acquired points, and on the surface reflection property (the reflection is stronger for brighter surfaces). Thus, the noise can be approximated with a Gaussian model [9], [3], [16] where the mean is 0 and the standard deviation is given by:

$$\sigma_z = \frac{c}{4\pi f_{mod}} \frac{\sqrt{I/2}}{A} \tag{1}$$

where, $z$ represents the depth, $f_{mod}$ is the modulation frequency of the signal sent by the ToF emitters, $A$ is the amplitude value at the considered pixel, $I$ is the corresponding intensity value and $c$ is the speed of light.

It is clear from (1) that the precision grows with the amplitude $A$ while it is inversely related to intensity $I$. Notice that the intensity depends on two factors: the received amplitude $A$ and the background illumination. When the amplitude of the received signal increases, the overall precision also increases (the dependency is with the square root w.r.t $I$ and linear w.r.t. $A$), while the background illumination affects only $I$ and reduces the precision.

The standard deviation of the error can be mapped into confidence values $\hat{C}_A$ for the point $p$ under examination. A simple approach [14] is to define two thresholds, $\sigma_{min}$ and $\sigma_{max}$, and then linearly map the values between them to the $[0, 1]$ interval as follows:

$$\hat{C}_A(p) = \begin{cases} 1 & if\ \sigma_z \leq \sigma_{min} \\ \frac{\sigma_{max} - \sigma_z}{\sigma_{max} - \sigma_{min}} & if\ \sigma_{min} < \sigma_z < \sigma_{max} \\ 0 & if\ \sigma_z \geq \sigma_{max} \end{cases} \tag{2}$$

### B. Confidence from depth variance

Another critical issue for ToF sensors is the mixed pixel effect [16], [3], occurring when a ToF sensor deals with points close to a depth edge. In this circumstance, the sensor can capture data relative to different surfaces at different distances estimating a depth value that is typically a convex combination of the two depth values. This issue leads to the fact that depth data is less reliable in the proximity of edges; this problem is amplified by the fact that current ToF sensors have a quite low resolution. A simple way to model this issue consists in using the local variance or the average absolute difference w.r.t. the considered point $p$ in the 8-neighborhood $\mathcal{N}(p)$ of $p$. A possible approach (ST-D$_T$, introduced in [14]) is to compute the confidence $\hat{C}_V$ by selecting a threshold $T_h$ and then linearly mapping the absolute differences from the $[0, T]$ interval into $[0, 1]$ as follows:

$$\hat{C}_V(p) = \max \left\{ 0;\ 1 - \frac{1}{T_h \cdot |\mathcal{N}(p)|} \sum_{j \in \mathcal{N}(p)} |z(p) - z(j)| \right\} \tag{3}$$

where $z(p)$ is the depth of the point $p$ under examination and $z(j)$ are the depth values within $\mathcal{N}(p)$.

Notice that also DA and DS measures exploit the idea that range image estimation is not reliable on depth discontinuities, but these work on the disparity domain and not in the depth one. Therefore, by converting the ToF depth in a disparity map, and by assigning each floating point disparity value $\hat{d}$ from the ToF sensor to a bin $\lfloor\hat{d}\rfloor : \lceil\hat{d}\rceil$, DA, DS and other features used by O1 (see Section III) can be easily adapted to ToF data enabling their deployment on this sensor as well.

### C. Confidence from machine learning approaches

An alternative strategy relies on machine learning approaches modeling confidence estimation, similarly to the stereo case reported in Section III.

However, a critical issue for this family of approaches is the lack of large datasets with ground truth needed for training. A possible workaround for this is to use synthetic data to train the ML algorithms [23], [44]. The method of [23] uses both amplitude and disparity information as input features and feeds them to a CNN in order to estimate confidence. The approach uses a CNN with 6 layers using ReLU activation functions. The first $5 \times 5$ layer extracts 64 features while the other layers use $3 \times 3$ kernels respectively producing 128, 128, 128, 256 and 2 features. There is no pooling operation in order to preserve the original resolution. The method of [23] introduces a confidence estimation strategy developed explicitly for ToF-stereo fusion feeding also stereo data to CNN (see Section V). As aforementioned, [23] uses a synthetic dataset built with a ToF simulator [45] in order to perform training. We followed the same training approach in this paper.

A more complex deep network with a coarse-fine structure has been exploited in [36] to estimate the noise of the ToF sensor. This work also uses features extracted from multi-frequency data to estimate the multi-path interference, a critical issue strongly reducing the confidence of ToF data. Even if the target of the work is the estimation of the noise, it can be mapped to a confidence value.

As highlighted before, some machine-learning strategies introduced for stereo matching can be applied seamlessly to ToF data as well. In particular, methods processing only depth information (i.e., O1 and CCNN) and eventually RGB image (i.e., LGC, when a side color image is available) are compatible for our purpose. In this paper, for the first time, we apply such strategies to ToF data as well, both studying how they behave when dealing with outliers detection and how they perform at fusion compared to prior proposals from literature. As for the strategies mentioned above [23], [36], since other datasets are not available, synthetic training samples are required in order to achieve good performance on both tasks. Finally, differently from [23], [36], we point out that these techniques minimize a binary-cross entropy loss. We will show with extensive experimental results, the impact of the two different loss functions on both outlier detection and fusion problems.

## V. JOINT ESTIMATION OF STEREO AND TOF CONFIDENCE

In order to combine stereo and ToF information, it is paramount to have two consistent ways of evaluating confidence. Even if the two confidence maps can be estimated separately, we will show in Section VII that the best results are achieved by jointly estimating the two maps.

The first work to introduce this idea was [23], where a CNN taking in input a multi-channel representation containing stereo and ToF data, projected on the reference camera of the stereo system, was in charge of estimating confidence scores for both sensors. This approach has been refined in [44]: we will refer to it as ST-CNN* in the results section. From now on the * is used to highlight methods jointly estimating confidence for both sensors together. In [44] the CNN takes in input multiple clues, i.e., the two disparity maps, the ToF amplitude information (used also by [14] but not by the other strategies considered in this work) and the difference between the left image of the stereo system and the right one warped over it using the stereo disparity. As shown in [44], the two additional channels allow to obtain a slight improvement in the confidence estimation w.r.t. using only disparity information. This network has two outputs and can be trained on synthetic data by minimizing a loss containing two components, one for each sensor. More in detail, the confidence is computed as a negative exponential function of the error, i.e., $C(p) = e^{-|\hat{d}(p)-d(p)|}$ and the loss to be minimized is:

$$\mathcal{L} = \sum \left(\hat{C}_T(p) - C_T(p)\right)^2 + \sum \left(\hat{C}_S(p) - C_S(p)\right)^2 \quad (4)$$

where $C_S(p)$ and $C_T(p)$ are the ground truth confidence values for stereo and ToF while $\hat{C}_S(p)$ and $\hat{C}_T(p)$ are the ones estimated from the network.

Consistently, to account for both sensors, the state-of-the-art CNN-based confidence measures CCNN [19] and LGC [43] originally proposed for stereo can be modified to jointly infer confidence scores for stereo and ToF by doubling the inputs and outputs of the network. In the next of this paper we will refer to their joint confidence estimation respectivelly with CCNN* and LGC*. Nonetheless, in contrast to ST-CNN* [25], their training relies only on depth data provided by the two sensors without any additional feature. Moreover, since approaches developed for stereo traditionally minimize a binary cross-entropy loss and, in this case, both modalities could be correct in terms of inlier vs outlier classification, we convert the task to a multi-labeling classification problem [46] and minimize for the following objective loss:

$$\begin{aligned}\mathcal{L} = &\sum \Big(C_T(p)\log\hat{C}_T(p) + (1 - C_T(p)))\log(1-\hat{C}_T(p)\Big) \\ &+ \sum \Big(C_S(p)\log\hat{C}_S(p) + (1 - C_S(p)))\log(1-\hat{C}_S(p)\Big)\end{aligned} \quad (5)$$

where in this case $C_T(p)$ and $C_S(p)$ are binary labels that can be 0 or 1 depending if the corresponding sensor has an error

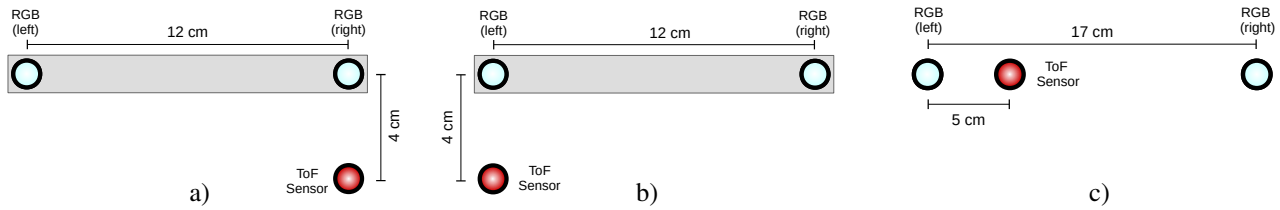


Fig. 1. Sensor arrangements: a) setup used in the SYNTH3 dataset; b) setup used in the REAL3 dataset; c) setup used in the LTTM5 dataset.

smaller than a pre-defined threshold, while $\hat{C}_T(p)$ and $\hat{C}_S(p)$ are the ones estimated from the network.

Concerning O1 [41], the joint training is not feasible without significant modifications, and hence, it has been trained to infer confidence estimation independently for stereo and ToF.

## VI. Fusion of stereo and ToF data

Data fusion is a widely adopted strategy in many application fields such as wireless sensor networks [47], remote sensing [48], and network traffic analysis [49] to name a few. Concerning depth sensor fusion, as previously pointed out in Sections III and IV, these two sensor technologies have rather complementary strengths and drawbacks making data fusion based on such setup quite appealing to obtain reliable depth information. Specifically, we will consider a trinocular setup like the ones depicted in Figure 1.

We will assume that the setup has been calibrated (we used the approach of [15] for this purpose) and that ToF data has been reprojected to the stereo viewpoint and interpolated to the same resolution of the stereo information. For the interpolation, we used the method of [34] based on an extended version of the cross bilateral filter.

In the considered setup, two different depth (or disparity) fields relating to the same viewpoint and at the same resolution are available. Different strategies can be exploited to combine the output of the two sensors taking into account confidence estimation. Purposely, we consider two simple approaches and a more advanced fusion strategy.

A first straightforward solution, referred to as *Highest Hypothesis (HH)*, consists of selecting at each pixel location, the disparity source (stereo or ToF) that has the highest confidence. Despite its simplicity, this strategy is fast, and, provided that confidence information is reliable, allows to significantly reduce artifacts when one of the two approaches is entirely unreliable (e.g., in case of wrong matches for the stereo system).

A second strategy, referred to as *Weighted Average (WA)*, consists of a weighted average of the two disparity values $d_T$ and $d_S$, computed according to the estimated confidence values as follows:

$$d = \frac{(\hat{C}_T + \epsilon) * d_T + (\hat{C}_S + \epsilon) * d_S}{\hat{C}_T + \hat{C}_S + 2\epsilon} \tag{6}$$

where $\epsilon$ is a small constant introduced to avoid issues when both acquisition systems have confidence close to $0$. Compared to the previous one, this strategy is more flexible and can output any depth value in the middle between the 2 measures. It typically yields better results when the two depth values are both reliable. However, if one sensor provides a wrong value and its confidence score is low but not close to $0$ can easily lead to artifacts. It is worth observing that although not very reliable, as reported next, an additional strategy, referred to as *Average*, can be obtained by neglecting the confidence contribution in WA (i.e., assuming the same weight for both sensors).

Finally, we consider a more advanced fusion strategy referred to as *LC*, based on the Locally Consistent fusion framework, introduced in [50] for stereo disparity refinement and extended to stereo-ToF fusion in [34], [14]. In its original formulation [50], the Locally Consistent framework aimed at inferring depth from a stereo pair exploiting a patch-based strategy and assuming piece-wise smooth surfaces in the sensed scene. Specifically, it analyzes the multiple depth hypotheses enforced for each point during the local processing in order to determine the most likely one accordingly. When tackling sensor fusion, the rationale behind this strategy can be exploited for reasoning about multiple depth maps, for instance, obtained by stereo and ToF as in [34]. Moreover, such an approach for sensor fusion can be further improved by taking into account confidence estimation as done in [14] and in the experiments reported in the next section.

## VII. Experimental Results

The experimental evaluation has been carried out on three different datasets, a synthetic dataset and two smaller sets of real-world scenes. Since this work proposes both a new set of confidence measures and various data fusion strategies, we divide the experimental evaluation into two parts: we firstly assess confidence estimation and then analyze the data fusion results according to standard evaluation protocols.

### A. Datasets

The first dataset is the SYNTH3 dataset [23]: it contains 55 synthetic scenes created using the Blender 3D rendering software and the Sony ToF Explorer simulator from Sony EUTec (that is based on the work of [45]). The parameters of the virtual cameras and their arrangement have been chosen in order to resemble an acquisition system composed by a Kinect v2 ToF sensor below a ZED stereo camera. Fig. 1a shows the camera setup: the stereo system has a baseline of 12 cm and the ToF sensor is placed below the right stereo camera at a distance of 4 cm. The data is split into a training set with 40 scenes and a test set with the remaining 15 scenes. The scenes have a large variability and include indoor and outdoor environments of very different sizes with objects of various

| | ST-CNN* | CCNN / CCNN* | LGC / LGC* |
|---|---|---|---|
| Learning rate | $10^{-1}$ | $10^{-3}$ | $10^{-3}$ |
| Epochs | 500 | 14 | 14/1600/14 |
| Optimizer | AdaDelta [51] | SGD | SGD |
| Regularization | $l_2$ ($\lambda = 10^{-2}$) | - | - |

TABLE I
ST-CNN, CCNN AND LGC TRAINING HYPER-PARAMETERS.

shapes, material and color. Notice that this is the only dataset large enough to perform training of ML-based approaches in a ToF-stereo fusion framework. Hence, all the learning-based confidence measures have been trained on this dataset.

The second dataset is the REAL3 dataset [44]; it contains 8 real-world scenes, and due to its small size has been used only for testing purposes. In contrast to the previous one, it contains real-world data. The scenes have been acquired with a Kinect v2 ToF sensor and a ZED stereo camera, deploying the SGM algorithm [42], while ground truth information has been obtained using two synchronized color cameras and a line laser (see [44] for more details on how the dataset has been created). The ToF and stereo camera placement is depicted in Fig. 1b, it is as in the previous dataset but with the ToF sensor below the left camera. The scenes are all indoor scenarios and include both simple flat surfaces and objects with a more complex geometry made of different materials.

The last dataset is LTTM5 [16]: it is a real-world dataset containing only 5 scenes all depicting various objects put on a table acquired with a stereo system and a ToF camera arranged as in Fig. 1c, i.e., with a larger baseline of 17 cm and the ToF sensor placed between the two color cameras (closer to the left one). Despite its small size, it is interesting since it has been used to evaluate many stereo-ToF fusion approaches and allows to compare with the state-of-the-art in the field.

### *B. Training of Learning-based Approaches*

Some of the stereo and ToF confidence estimators employed in this paper rely on machine learning techniques. In particular, the deep learning approaches ST-CNN, CCNN and LGC have been trained using the training split of the SYNTH3 dataset and the hyper-parameters shown in Table I. Please notice that the different methods have different parameters, but the networks jointly estimating ToF and stereo confidences (ST-CNN*, CCNN* and LGC*) share the same hyper-parameters of their base implementation. For what concerns the LGC method, CCNN, ConfNet and the final module have been trained for 14, 1600 and 14 epochs, respectively.

### *C. Confidence Evaluation*

We start from evaluating confidence measures on stereo and ToF data according to the standard protocol used in this field [7], [18] on the 3 datasets. Tables II, III and IV show the AUC values of the different considered confidence metrics for both stereo and ToF with the error threshold set to 1, 2 and 4 respectively. All tables report, in different columns, results on the three datasets mentioned above. On the bottom, we also report both the optimal AUC obtained according to [7], [18]. The scores have been multiplied by a factor $10^2$ to ease readability.

| | SYNTH3 | | REAL3 | | LTTM5 | |
|---|---|---|---|---|---|---|
| | Stereo | ToF | Stereo | ToF | Stereo | ToF |
| ST-D | 12.97 | 14.71 | 53.53 | 79.35 | 10.00 | 31.56 |
| DA | 4.95 | 12.68 | 44.52 | 72.15 | 4.29 | 29.25 |
| DS | 5.46 | 18.05 | 45.46 | 70.30 | 4.67 | 30.53 |
| O1 | 4.10 | 10.45 | 42.91 | 72.08 | 3.85 | 22.81 |
| ST-CNN* | 3.26 | 11.05 | 45.02 | **66.85** | 4.47 | 22.16 |
| CCNN | 5.24 | 20.63 | 44.35 | 69.94 | 3.20 | 20.84 |
| CCNN* | **2.59** | **10.41** | **40.19** | 75.10 | 2.84 | **15.28** |
| LGC | 3.34 | 16.40 | 43.88 | 67.33 | 3.13 | 18.82 |
| LGC* | 2.75 | 12.08 | 41.03 | 76.41 | **2.25** | 18.50 |
| Opt.AUC | 1.54 | 3.84 | 34.96 | 48.09 | 0.77 | 8.76 |
| Err.rate (%) | 15.16 | 21.10 | 67.04 | 78.11 | 12.09 | 37.08 |

TABLE II
CONFIDENCE EVALUATION: AUC VALUES ($\times 10^2$) WITH THRESHOLD 1.

| | SYNTH3 | | REAL3 | | LTTM5 | |
|---|---|---|---|---|---|---|
| | Stereo | ToF | Stereo | ToF | Stereo | ToF |
| ST-D | 9.96 | 5.26 | 47.18 | 51.91 | 6.72 | 14.13 |
| DA | 3.46 | 4.69 | 37.92 | 42.19 | 2.69 | 11.67 |
| DS | 3.87 | 11.02 | 38.82 | 39.34 | 3.00 | 14.84 |
| O1 | 2.60 | 2.39 | 36.12 | 41.61 | 2.33 | 7.24 |
| ST-CNN* | 1.95 | 7.12 | 35.25 | 36.00 | 2.06 | 4.67 |
| CCNN | 3.56 | 8.96 | 37.74 | 37.20 | 2.04 | 5.73 |
| CCNN* | 1.38 | 2.21 | **30.63** | 44.22 | 0.90 | **3.50** |
| LGC | 2.28 | 6.00 | 37.47 | **30.79** | 1.91 | 4.88 |
| LGC* | **1.35** | **2.20** | 31.13 | 44.12 | **0.73** | 4.18 |
| Opt.AUC | 0.94 | 0.82 | 26.91 | 18.22 | 0.39 | 2.06 |
| Err.rate (%) | 11.85 | 11.91 | 60.24 | 49.31 | 8.60 | 18.52 |

TABLE III
CONFIDENCE EVALUATION: AUC VALUES ($\times 10^2$) WITH THRESHOLD 2.

Starting from the SYNTH3 dataset, it is possible to see how learning-based approaches (O1 [41], CCNN [19], LGC [43], ST-CNN* [44]) have in general better performance than traditional ones (ST-D, DA and DS although these latter two methods are rather effective). Here, we refer to the joint application of the stereo and ToF confidences ST-D$_S$ and ST-D$_T$ with ST-D as introduced in Sections III and IV.

Focusing on CCNN and LGC approaches: they have been trained both independently on stereo and ToF and jointly on the two sensors (tagged in this case, respectively, as LGC* and CCNN* in the tables). ST-CNN* has instead always been trained jointly as initially proposed in [44]. We can note that the joint training on the stereo and ToF data consistently yields much better performance. Moreover, as reported later, such a strategy will be particularly helpful when dealing with the fusion problem where the consistency between the two confidence metrics is a fundamental requirement for achieving high performance.

According to tables II, III and IV, on the SYNTH3 dataset the two best performing approaches, for both stereo and ToF, are CCNN* and LGC* jointly trained on both modalities. However, it is worth noticing that these two approaches are trained to minimize a classification loss function that is ideal for reducing the AUC, while ST-CNN* is trained with a regression loss where the ground truth confidence measure has been computed as a function of the sensor disparity error thorugh Equation (4). The CCNN* approach is the best when the AUC threshold is set to 1 while LGC* leads to better performance on both stereo and ToF data when considering larger thresholds. Thus, LGC is less effective with smaller errors but better when dealing with higher magnitude outliers. In particular, in these latter cases (Tables III and IV) LGC gets quite close to the optimal AUC. The ST-CNN* approach has a relatively good performance on this dataset, especially for what concerns the stereo data, demonstrating again that

| | SYNTH3 | | REAL3 | | LTTM5 | |
|---|---|---|---|---|---|---|
| | Stereo | ToF | Stereo | ToF | Stereo | ToF |
| ST-D | 7.79 | 3.58 | 40.27 | 3.95 | 5.67 | 3.68 |
| DA | 2.35 | 3.56 | 32.10 | 3.35 | 2.39 | 7.22 |
| DS | 2.76 | 10.01 | 32.75 | 3.25 | 2.68 | 11.69 |
| O1 | 1.59 | 1.56 | 29.89 | 3.44 | 2.04 | 1.84 |
| ST-CNN* | 1.31 | 6.09 | 25.92 | 3.63 | 1.58 | 2.11 |
| CCNN | 2.15 | 6.43 | 31.65 | 3.85 | 1.78 | 1.74 |
| CCNN* | 0.86 | 1.48 | **21.64** | 3.60 | 0.62 | 1.23 |
| LGC | 1.47 | 4.22 | 31.65 | 2.79 | 1.63 | 1.51 |
| LGC* | **0.78** | **1.41** | 21.88 | **2.72** | **0.51** | **1.08** |
| Opt.AUC | 0.58 | 0.45 | 19.05 | 0.54 | 0.30 | 0.64 |
| Err.rate (%) | 9.36 | 9.01 | 51.87 | 7.43 | 7.41 | 11.03 |

TABLE IV
CONFIDENCE EVALUATION: AUC VALUES ($\times 10^2$) WITH THRESHOLD 4.

| | SYNTH3 | | | REAL3 | | | LTTM5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | LC | WA | HH | LC | WA | HH | LC | WA | HH |
| ST-D | 2.02 | 2.13 | 2.59 | 2.97 | 3.04 | 3.33 | 2.79 | 2.78 | 3.04 |
| DA | 2.40 | 2.01 | 2.33 | 3.91 | 5.12 | 3.66 | 3.23 | 2.91 | 3.58 |
| DS | 2.39 | 2.15 | 2.46 | 3.96 | 5.42 | 4.99 | 3.46 | 3.36 | 3.94 |
| O1 | 1.99 | 1.77 | 1.98 | 3.23 | 4.60 | 3.90 | 3.25 | 2.97 | 3.49 |
| ST-CNN* | 1.97 | **1.66** | **1.80** | 2.88 | 4.05 | 3.78 | **2.70** | **2.70** | 3.12 |
| CCNN | 2.04 | 1.80 | 2.03 | 3.24 | 4.00 | 3.25 | 3.44 | 2.90 | 3.42 |
| CCNN* | 1.92 | 1.91 | 2.03 | **2.40** | 2.75 | **2.60** | 2.74 | 2.91 | 3.00 |
| LGC | 2.00 | 1.74 | 1.95 | 3.21 | 3.77 | 2.87 | 3.53 | 3.13 | 3.53 |
| LGC* | **1.83** | 1.89 | 2.03 | 2.49 | **2.65** | **2.60** | 2.75 | 2.85 | **2.95** |
| Average | | 2.34 | | | 7.50 | | | 3.04 | |
| Stereo | | 3.67 | | | 14.19 | | | 4.47 | |
| ToF | | 2.18 | | | 3.28 | | | 3.40 | |

TABLE V
FUSION ACCURACY MEASURED WITH RMSE.

| **SYNTH3** | $\epsilon = 1$ | | | $\epsilon = 2$ | | | $\epsilon = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | LC | WA | HH | LC | WA | HH | LC | WA | HH |
| ST-D | 14.05 | 13.96 | 13.57 | 4.93 | 4.99 | 4.72 | 2.26 | 2.23 | 2.24 |
| DA | 9.89 | 8.95 | 7.58 | 4.74 | 4.57 | 4.50 | 2.54 | 2.41 | 2.47 |
| DS | 9.88 | 8.94 | 8.49 | 4.79 | 4.72 | 4.41 | 2.57 | 2.57 | 2.47 |
| O1 | 9.98 | 8.94 | 8.86 | 4.35 | 4.19 | 4.08 | 2.25 | 2.12 | 2.13 |
| ST-CNN* | **8.26** | 8.13 | 6.88 | **3.77** | 4.02 | **3.50** | 1.94 | 1.81 | 1.72 |
| CCNN | 9.79 | 9.35 | 10.36 | 4.63 | 4.69 | 4.03 | 2.40 | 2.30 | 1.72 |
| CCNN* | 9.61 | 8.94 | 7.54 | 4.20 | 3.52 | 3.69 | 1.94 | 1.71 | 1.72 |
| LGC | 9.84 | 9.62 | **6.69** | 4.67 | 4.79 | 3.74 | 2.40 | 2.32 | 1.81 |
| LGC* | 9.35 | **7.78** | 9.69 | 3.92 | **3.15** | 3.75 | **1.71** | **1.65** | **1.65** |
| Average | | 11.30 | | | 6.30 | | | 3.42 | |
| Stereo | | 10.12 | | | 6.80 | | | 4.22 | |
| ToF | | 14.92 | | | 5.35 | | | 2.25 | |

TABLE VI
FUSION ACCURACY MEASURED AS THE *PERCENTAGE OF WRONG PIXELS* ON THE SYNTH3 DATASET.

learning-based approaches jointly trained on stereo and ToF are the best family of solutions. Although O1 performs relatively well, it is always outperformed by LGC* and CCNN* with all thresholds.

A fundamental problem for deep learning approaches is the risk of focusing too much on the training dataset, that in this case, for the reasons outlined before, is entirely composed of synthetic data. Therefore, it is essential to assess how they can generalize to real-world scenarios represented by datasets REAL3 and LTTM5. The tests on the REAL3 dataset show how the learned confidence measure keep excellent performance even if the gap with traditional ones gets smaller compared to the synthetic dataset.

On the REAL3 dataset, as on synthetic data, the competition is still between LGC* and CCNN* jointly trained on the stereo and ToF data. Nonetheless, considering the smallest threshold, ST-CNN* turns out to be the best on ToF data. Moreover, the gap between traditional DA and DS confidence measures and learning-based ones is reduced if compared to SYNTH3.

Finally, on the other real-world LTTM5 dataset LGC* and CCNN* approaches are overall the best ones. In particular, LGC* is always the best on stereo data and with the ToF sensor with threshold 4. CCNN* is the best in the other 2 cases with ToF data. ST-CNN* and O1 have overall good performance and again DA and DS are less reliable but with a smaller gap compared to the synthetic case.

From this exhaustive evaluation we can notice how results are consistent on all the experiments, showing that learning approaches jointly trained for estimation of ToF and stereo confidences are the best solution and that, on average, LGC* is the best technique.

### *D. Fusion of Stereo and ToF data*

Once assessed the performance of confidence measures, we leverage this cue for the fusion of the two disparity fields generated by stereo and ToF sensors. Specifically, we evaluated all the confidence measures considered in Sections III, IV and V with the fusion strategies outlined in Section VI. The outcome concerning the Root Mean Square Error (RMSE) between the fused disparity maps and ground truth data is shown in Table V. We also report in the table the RMSE for raw ToF and stereo data and the simple fusion scheme averaging the two disparity values at each location.

On the SYNTH3 dataset, the ST-CNN* approach allows obtaining excellent results deploying simple fusion strategies, especially with WA. On the other hand, with LC, the LGC* approach performs better. This behavior changes on the REAL3 dataset where LC achieves the best results with CCNN* although straightforward fusion strategies coupled with LGC* and CCNN* do pretty well. A similar trend is observed on the LTTM5 dataset, with LC and WA coupled with ST-CNN* yielding the best results although CCNN* and LGC* allow to obtain slightly worse accuracy. A first thing to notice from the table, and experiments on other datasets reported next, is that the simple average of the two disparity fields is often unable to improve the accuracy of the output disparity map compared to the raw depth maps coming from the ToF and stereo devices, resulting in between the two in terms of accuracy. Moreover, the use of confidence maps in the fusion process helps to reliably combine the two depth data sources. By using accurate confidence maps, provided by learning-based methods, and very accurate data like the synthetic one available in SYNTH3 even the straightforward fusion strategies WA and HH allow getting excellent results. However, the more complex LC fusion strategy turns out quite useful on real-world datasets (especially with REAL3). Moreover, a second thing to notice is that training ST-CNN* by minimizing a regression loss turns out not optimal concerning AUC evaluation, but it helps to obtain a better granularity when dealing with small errors in the LC framework. ST-CNN* however also exploits additional features that help in driving the fusion process, as pointed out in Section IV.

Tables VI, VII and VIII report for the three datasets the fusion outcome in term of percentage of wrong pixels. The error bound set to discriminate the goodness of the disparity estimation is respectively set to 1, 2 and 4 pixels as for confidence evaluation.

Concerning the evaluation on the SYNTH3 dataset, from

| REAL3 | $\epsilon = 1$ | | | $\epsilon = 2$ | | | $\epsilon = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | LC | WA | HH | LC | WA | HH | LC | WA | HH |
| ST-D | 74.07 | 81.77 | 81.58 | 47.11 | 49.24 | 53.13 | 6.04 | 6.49 | 7.06 |
| DA | 57.93 | 70.41 | 51.16 | 33.11 | 38.76 | 27.98 | 3.68 | 22.41 | **4.17** |
| DS | 55.56 | 66.53 | 51.61 | 31.27 | 38.27 | 28.90 | 3.77 | 23.63 | 6.88 |
| O1 | 57.89 | 69.53 | 55.91 | 33.06 | 39.91 | 32.27 | 3.94 | 22.13 | 6.68 |
| ST-CNN* | **54.75** | **61.06** | 51.22 | **30.35** | 37.53 | 30.03 | 3.93 | 16.89 | 8.45 |
| CCNN | 60.13 | 77.00 | 76.42 | 35.52 | 39.10 | 49.83 | 4.70 | 17.74 | 6.64 |
| CCNN* | 57.55 | 69.63 | **48.35** | 32.54 | 29.26 | **26.63** | **3.11** | 6.79 | 5.46 |
| LGC | 60.49 | 77.24 | 58.71 | 36.00 | 38.97 | 35.26 | 5.26 | 15.21 | 6.13 |
| LGC* | 57.14 | 66.04 | 48.58 | 32.15 | **27.73** | 26.96 | 3.25 | **5.65** | 5.50 |
| Average | 81.33 | | | 50.12 | | | 34.17 | | |
| Stereo | 57.44 | | | 50.13 | | | 42.17 | | |
| ToF | 83.67 | | | 55.37 | | | 7.93 | | |

TABLE VII
FUSION ACCURACY MEASURED AS THE *PERCENTAGE OF WRONG PIXELS* ON THE REAL3 DATASET.

| LTTM5 | $\epsilon = 1$ | | | $\epsilon = 2$ | | | $\epsilon = 4$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | LC | WA | HH | LC | WA | HH | LC | WA | HH |
| ST-D | 28.94 | 33.56 | 33.85 | 10.69 | 13.74 | 14.78 | 3.87 | 5.18 | 5.03 |
| DA | 18.36 | 18.07 | 13.54 | 6.84 | 8.73 | 7.32 | 4.53 | 5.62 | 5.27 |
| DS | **15.13** | **14.70** | 13.37 | 6.56 | 9.00 | 7.72 | 5.09 | 7.01 | 6.12 |
| O1 | 16.55 | 19.49 | 16.46 | 6.88 | 9.55 | 8.53 | 4.59 | 6.08 | 5.20 |
| ST-CNN* | 17.24 | 20.68 | 20.05 | **5.40** | 9.11 | 7.50 | 3.59 | 6.01 | 5.24 |
| CCNN | 19.94 | 23.69 | 31.01 | 7.77 | 10.84 | 13.50 | 4.96 | 6.64 | 5.50 |
| CCNN* | 17.54 | 18.07 | 18.04 | 5.82 | **6.57** | 5.99 | 3.54 | 4.59 | 4.34 |
| LGC | 20.64 | 25.17 | 23.02 | 8.01 | 10.15 | 10.56 | 5.19 | 6.62 | 5.95 |
| LGC* | 17.48 | 18.28 | **10.79** | 5.75 | 6.77 | **5.70** | **3.53** | **4.58** | **4.28** |
| Average | 24.03 | | | 12.46 | | | 8.85 | | |
| Stereo | 13.72 | | | 9.31 | | | 7.67 | | |
| ToF | 37.14 | | | 17.27 | | | 6.66 | | |

TABLE VIII
FUSION ACCURACY MEASURED AS THE *PERCENTAGE OF WRONG PIXELS* ON THE LTTM5 DATASET.

Table VI we can notice that the trend is substantially the same observed when evaluating the RMSE. The simple fusion strategies WA and HH yield the best performance when coupled with LGC with threshold 1 and with LGC* with the other thresholds. On the REAL3 dataset (Table VII), HH with CCNN* is the best fusion method regarding the percentage of pixels with error less than 2 pxl, instead LC in combination with the same confidence measure is able to reduce the most the error higher than 4 pxl. In all cases, CCNN* is the confidence measure yielding the best results. Regarding the LTTM5 dataset (Table VIII), results are more variegate: LC achieves the lowest percentage of bad pixels with threshold 2 and 4, respectively coupled with ST-CNN* and CCNN*, while HH confirms to be the best method with threshold 1, especially when coupled with LGC*.

From the results reported so far, we can notice that the confidence measure with the best AUC is not necessarily the best method for depth fusion. On the other hand, it is quite evident that the best results are typically obtained by confidence measures showing good performance according to the AUC metric. Moreover, the joint training of confidence measures turns out to be very useful as done for ST-CNN*, CCNN* and LGC*.

In Table IX we also report the evaluation on the LTTM5 dataset of 4 state-of-the-art approaches: namely, [34] that was the first to use LC for stereo-ToF fusion, the MAP-MRF Bayesian frameworks proposed in [28] and [16], and the approach based on bilateral filtering of the cost volume introduced in [52]. The results clearly show how the best strategies proposed in this paper outperform previous approaches known in the literature.

| | LC ST-CNN* | WA ST-CNN* | HH LGC* | [34] | [52] | [28] | [16] |
|---|---|---|---|---|---|---|---|
| RMSE | **2.70** | **2.70** | 2.95 | 3.17 | 3.31 | 3.34 | 3.49 |

TABLE IX
COMPARISON WITH STATE-OF-THE-ART FUSION METHODS ON THE SYNTH3 DATASET.

| | ST-D | DA | DS | O1 | ST-CNN* | CCNN* | LGC* |
|---|---|---|---|---|---|---|---|
| Runtime [ms] | 480.6 | 868.9 | 868.9 | 3534.2 | 102.6 | 26.7 | 131.9 |

TABLE X
RUNTIME OF THE CONFIDENCE ESTIMATION TECHNIQUES.

### *E. Qualitative Results*

Figures 2, 3 and 4 report qualitative experimental results on a sample scene extracted respectively from the SYNTH3, REAL3 and LTTM5 datasets. The first row contains the disparity maps coming from the ToF and stereo sensors and the results of the LC fusion using the ST-CNN* and the LGC* confidences. We selected these fusion strategies since they are on average the best performing on the 3 datasets. The second row contains the disparity error maps computed as the true disparity minus the target disparity. The color map encodes the correct estimation with green, the disparity overestimation with colder colors and disparity underestimation with warmer colors. From the figures, it is possible to notice well known issues of ToF and stereo. Specifically, concerning stereo, we can observe poor performance on textureless regions and in case of repeating patterns like the green box with the white grid in Fig. 4. On the other hand, the main issues of the ToF sensor arise on the object sides, due to the original low spatial resolution of the sensor, and the disparity underestimation near to corners, due to the multi-path interference. The error maps related to the two fusion strategies show a large overall reduction of the error: the proposed approaches are able to take the best from the 2 depth sources thus avoiding the stereo artifacts on un-textured regions and greatly reducing the MPI corruption in the ToF data. The third and fourth column contain the ST-CNN* and LGC* confidence maps which have values in the range from 0 (not reliable pixels) to 1 (highly reliable pixels). Both of of them follow quite accurately the error distribution, but with different behaviour. LGC* is more binarized, since it is trained using a classification loss, instead ST-CNN* has a smoother transition, since it is trained with a regression loss.

### *F. Runtime analysis*

In order to asses the computational complexity of various fusion methods considered, we report the runtime of all the steps of the various fusion methods on the REAL3 dataset. These tests have been carried out on a PC with an Intel i7-4790 CPU and an NVIDIA Titan X GPU (used only for CNN-based techniques). Starting from pre-processing steps, the stereo disparity map computation with SGM [42] takes 1.49 s and the reprojection and interpolation of the ToF depth on the reference camera of the stereo system takes 4.47 s. These two operations are carried out for all the fusion techniques.

Table X collects the runtime for confidence estimation. They range from just 26 ms for CCNN* to more than 3 s for O1. For

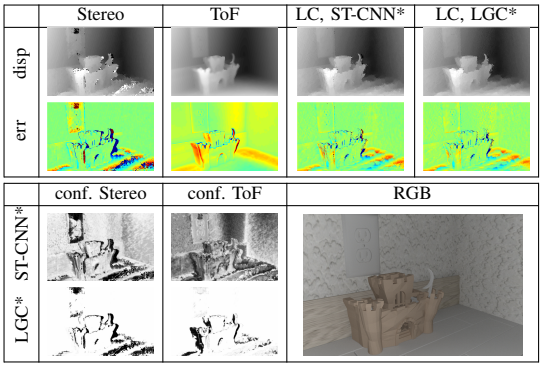


Fig. 2. Qualitative results on the SYNTH3 dataset.

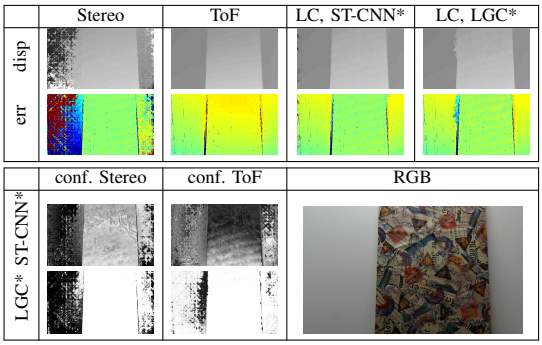


Fig. 3. Qualitative results on the REAL3 dataset.

methods allowing the joint confidence estimation of ToF and stereo, the separate estimation of the two confidences roughly doubles the computation time compared to the joint estimation since two inferences on two different networks are required.

Concerning the final fusion step, the simplicity of the average, HH and WA schemes allows us to perform these tasks in less than 5 ms. In contrast, LC is much more complex and requires about 35 s.

## VIII. Conclusions

Time-of-Flight and stereo are two popular depth sensing technologies with quite complementary strengths and limitations. For this reason, they are often combined to infer more accurate depth maps. Therefore, inspired by recent advances in stereo confidence estimation in this paper we introduce and evaluate learning-based confidence estimation strategies suited for depth data generated by ToF and stereo sensors showing how a joint training of such methods yields in general better performance. Moreover, deploying three fusion frameworks, we report how confidence estimation can effectively guide the fusion of data generated by the two depth sensing technology. Exhaustive experimental results show how the accurate confidence cues obtained allow to outperform state-of-the-art data fusion schemes even deploying straightforward fusion strategies. Future work will focus on the extension of the considered fusion strategies to video sequences, exploiting Recurrent Neural Networks (RNN) or dynamic system modeling strategies like the Kalman filter. Moreover, we will also consider the exploitation of an end-to-end deep learning approach for joint confidence estimation and depth fusion.

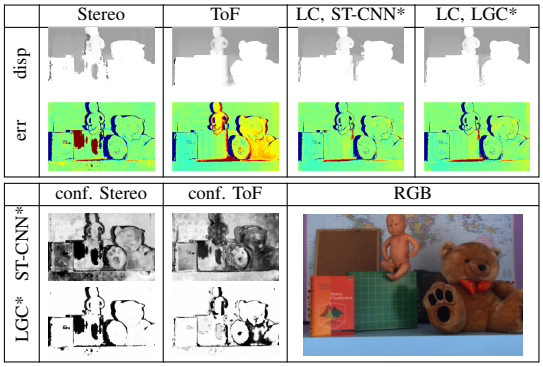


Fig. 4. Qualitative results on the LTTM5 dataset.

## References

[1] R. Nair, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, and D. Kondermann, "High accuracy tof and stereo sensor fusion at interactive rates," in *European Conference on Computer Vision*. Springer, 2012, pp. 1–11.

[2] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt, "Sra: Fast removal of general multipath for tof sensors," in *European Conference on Computer Vision*. Springer, 2014, pp. 234–249.

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

[4] "Middlebury stereo vision benchmark," http://vision.middlebury.edu/stereo/.

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

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

[7] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

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

[9] F. Remondino and D. Stoppa, Eds., *TOF Range-Imaging Cameras*. Springer, 2013.

[10] D. Piatti and F. Rinaudo, "Sr-4000 and camcube3.0 time of flight (tof) cameras: Tests and comparison," *Remote Sensing*, vol. 4, no. 4, pp. 1069–1089, 2012.

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

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

[13] G. Agresti and P. Zanuttigh, "Combination of spatially-modulated tof and structured light for mpi-free depth estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

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

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

[16] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, "Probabilistic ToF and stereo data fusion based on mixed pixels measurement models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2260–2272, 2015.

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

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

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

[20] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.
[21] M. Poggi and 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.
[22] M. Reynolds, J. Dobo, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 945–952.
[23] G. Agresti, L. Minto, G. Marin, and P. Zanuttigh, "Deep learning for confidence information in stereo and tof data fusion," in *ICCV Workshop: 3D Reconstruction meets Semantics*, Oct 2017.
[24] G. Agresti and Z. P., "Deep learning for multi-path error removal in tof sensors," in *Geometry Meets Deep Learning ECCV Workshop*, 2018.
[25] G. Agresti, H. Schaefer, P. Sartor, and P. Zanuttigh, "Unsupervised domain adaptation for tof data denoising with adversarial learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[26] J. Marco, Q. Hernandez, A. Muñoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez, "Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 219:1–219:12, nov 2017.
[27] R. Nair, K. Ruhl, F. Lenzen, S. Meister, H. Schäfer, C. Garbe, M. Eisemann, M. Magnor, and D. Kondermann, "A survey on time-of-flight stereo fusion," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 8200, pp. 105–127.
[28] J. Zhu, L. Wang, R. Yang, and 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.
[29] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 899–909, 2010.
[30] J. Zhu, L. Wang, R. Yang, J. E. Davis, and 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*, vol. 33, no. 7, pp. 1400–1414, 2011.
[31] B. Chen, C. Jung, and 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.
[32] ——, "Variational fusion of time-of-flight and stereo data for depth estimation using edge selective joint filtering," *IEEE Transactions on Multimedia*, pp. 1–1, 2018.
[33] G. Evangelidis, M. Hansard, and R. Horaud, "Fusion of Range and Stereo Data for High-Resolution Scene-Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2178 – 2192, 2015.
[34] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and 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.
[35] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and . Cuesta, "Pixel weighted average strategy for depth sensor data fusion," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2805–2808.
[36] G. Agresti and P. Zanuttigh, "Deep learning for multi-path error removal in ToF sensors," in *Geometry Meets Deep Learning ECCV Workshop*, 2018.
[37] G. Riegler, A. O. Ulusoy, H. Bischof, and A. Geiger, "Octnetfusion: Learning depth fusion from data," in *International Conference on 3D Vision (3DV)*, 2017, pp. 57–66.
[38] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
[39] A. Spyropoulos, N. Komodakis, and P. Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching." in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
[40] M. G. Park and K. J. Yoon, "Leveraging stereo matching with learning-based confidence measures," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
[41] M. Poggi and S. Mattoccia, "Learning a general-purpose confidence measure based on o(1) features and a smarter aggregation strategy for semi global matching," in *Proceedings of the 4th International Conference on 3D Vision, 3DV*, 2016.
[42] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.
[43] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proceedings of European Conference on Computer Vision (ECCV)*, September 2018.
[44] G. Agresti, L. Minto, G. Marin, and P. Zanuttigh, "Stereo and tof data fusion by learning from synthetic data," *Information Fusion*, vol. 49, pp. 161 – 173, 2019.
[45] S. Meister, R. Nair, and D. Kondermann, "Simulation of Time-of-Flight Sensors using Global Illumination," in *Vision, Modeling and Visualization*. The Eurographics Association, 2013.
[46] M. Poggi and S. Mattoccia, "Deep stereo fusion: combining multiple disparity hypotheses with deep-learning," in *Proceedings of the 4th International Conference on 3D Vision, 3DV*, 2016.
[47] D. Ciuonzo, G. Romano, and P. S. Rossi, "Channel-aware decision fusion in distributed mimo wireless sensor networks: Decode-and-fuse vs. decode-then-fuse," *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 2976–2985, 2012.
[48] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1253–1257, 2017.
[49] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Multi-classification approaches for classifying mobile app traffic," *Journal of Network and Computer Applications*, vol. 103, pp. 131–145, 2018.
[50] S. Mattoccia, "A locally global approach to stereo correspondence," in *Proc. of 3D Digital Imaging and Modeling (3DIM)*, October 2009.
[51] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
[52] Q. Yang, R. Yang, J. Davis, and 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.

**Matteo Poggi** received Master degree in Computer Science and PhD degree in Computer Science and Engineering from University of Bologna in 2014 and 2018 respectively. Currently, he is a Post-doc researcher at Department of Computer Science and Engineering, University of Bologna. His research interests include deep learning for depth estimation and embedded computer vision.

**Gianluca Agresti** received the Master degree in Telecommunication Engineering from University of Padova in 2016. Currently, he is a PhD student at the Department of Information Engineering of University of Padova. His research focuses on deep learning for ToF sensor data processing and multiple sensor fusion for 3D acquisition.

**Fabio Tosi** received the Master degree in Computer Science and Engineering at Alma Mater Studiorum, University of Bologna in 2017. He is currently in the PhD program in Computer Science and Engineering of University of Bologna, where he conducts research in deep learning and depth sensing related topics.

**Pietro Zanuttigh** received a Master degree in Computer Engineering at the University of Padova in 2003 where he also got the Ph.D. degree in 2007. Currently he is an assistant professor at the Department of Information Engineering. His research activity focuses on 3D data processing, in particular ToF sensors data processing, multiple sensors fusion for 3D acquisition, semantic segmentation and hand gesture recognition.

**Stefano Mattoccia** received a Ph.D. degree in Computer Science Engineering from the University of Bologna in 2002. Currently he is an associate professor at the Department of Computer Science and Engineering of the University of Bologna. His research interest is mainly focused on computer vision, depth perception from images, deep learning and embedded computer vision. In these fields, he has authored about 100 scientific publications and 3 patents.