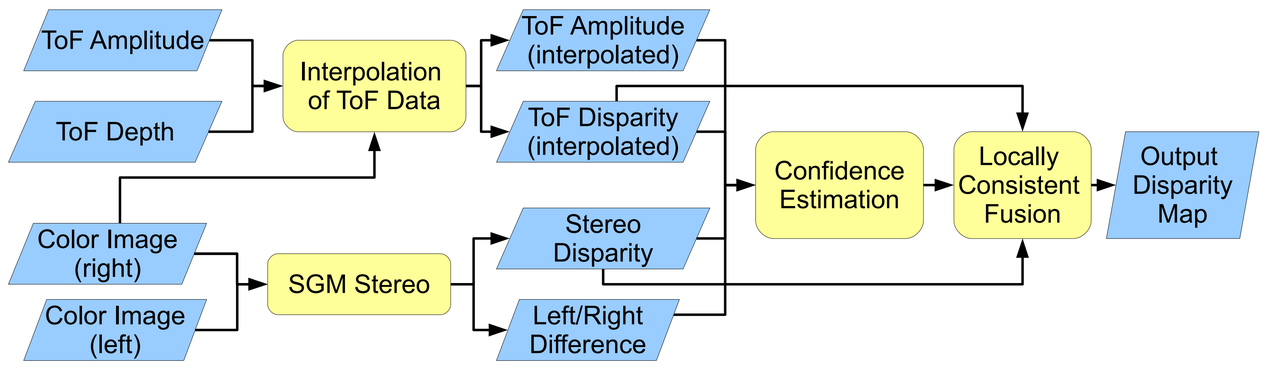
# Estimating Confidence Maps for ToF-Stereo Fusion Using Deep Learning

Gianluca Agresti, Ludovico Minto, Giulio Marin, Pietro Zanuttigh. University of Padova, Italy
{agrestig, mintolud, maringiu, zanuttigh}@dei.unipd.it

## Abstract

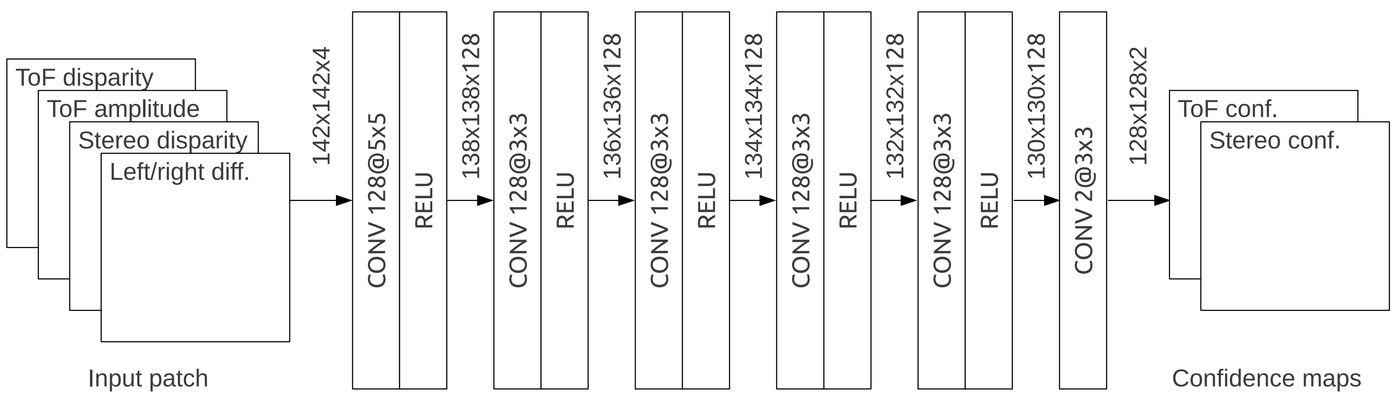
This paper proposes a novel framework for the fusion of depth data produced by a Time-of-Flight (ToF) camera and a stereo vision system. The key problem of balancing between the two sources of information is solved by extracting confidence maps for both sources using deep learning. We introduce a novel synthetic dataset accurately representing the data acquired by the proposed setup and use it to train a Convolutional Neural Network architecture. The machine learning framework estimates the reliability of both data sources at each pixel location. The two depth fields are finally fused enforcing the local consistency of depth data taking into account the confidence information. Experimental results show that the proposed approach increases the accuracy of the depth estimation.

## Proposed Method




## Confidence Estimation with CNN

The interpolated ToF amplitude and disparity, Stereo disparity and Left/Right difference are fed to a CNN trained for the confidence estimation for both Stereo and ToF.



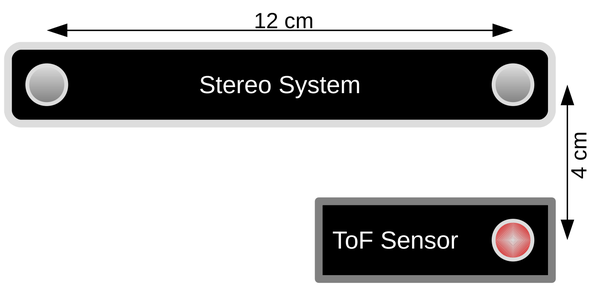

## Locally Consistent Fusion

For each pixel $f$ of the reference camera in the stereo vision system and for each disparity value $d$, a measure of the plausability is computed for the stereo disparity estimation, $\mathcal{P}_{f,g,S}(d)$, and for the ToF estimation, $\mathcal{P}_{f,g,T}(d)$. The plausability is function of the color and spatial consistency of the data in a certain region $\mathcal{A}$ surrounding $f$ [2]. The final disparity is evaluated by maximizing

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

where $P_T(g)$, $P_S(g)$ are the confidence maps computed with the proposed CNN.

## Stereo-ToF Acquisition System

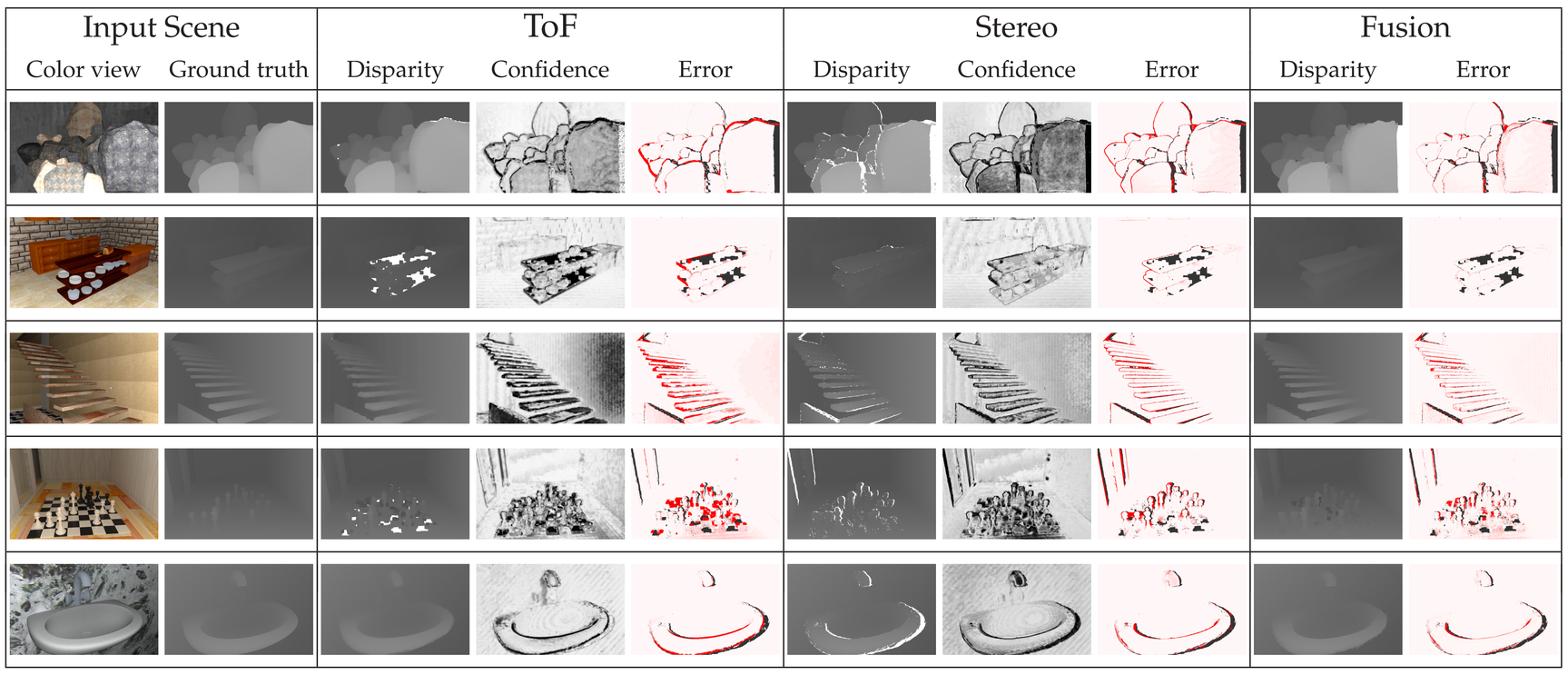



| | Stereo setup | ToF camera |
|---|---|---|
| Resolution | $1920 \times 1080$ | $512 \times 424$ |
| Horizontal FOV | $69°$ | $70°$ |
| Focal length | $3.2mm$ | $3.66mm$ |
| Pixel size | $2.2\mu m$ | $10\mu m$ |

## Synthetic Dataset

We introduce a synthetic dataset that is publicly available at [3]. We used it to train the CNN and evaluate the performance of the proposed method.

## Experimental Results




## Comparison

The table shows the average Root Mean Square Error (RMSE) evaluated on the proposed synthetic test set.

| Method | RMSE |
|---|---|
| Interpolated ToF | 2.19 |
| SGM Stereo | 3.73 |
| **Stereo-ToF Fusion** [1] | **2.06** |
| Marin et Al. [2] | 2.07 |

[1] Gianluca Agresti, Ludovico Minto, Giulio Marin and Pietro Zanuttigh. *"Deep Learning for Confidence Information in Stereo and ToF Data Fusion."* International Conference on Computer Vision Workshop 2017.
[2] Giulio Marin, Pietro Zanuttigh and Stefano Mattoccia. *"Reliable Fusion of ToF and Stereo Depth Driven by Confidence Measures."* European Conference on Computer Vision. Springer, 2016.
[3] `http://lttm.dei.unipd.it/paper_data/deepfusion/`