Learning confidence measures in the wild

Fabio Tosi
http://vision.disi.unibo.it/~ftosi
Matteo Poggi
http://vision.disi.unibo.it/~mpoggi
Alessio Tonioni
alessio.tonioni@unibo.it
Luigi Di Stefano
luigi.distefano@unibo.it
Stefano Mattoccia
http://vision.disi.unibo.it/~smatt

University of Bologna
Department of Computer Science and Engineering
Bologna, Italy

Abstract

Confidence measures for stereo earned increasing popularity in most recent works concerning stereo, being effectively deployed to improve its accuracy. While most measures are obtained by processing cues from the cost volume, top-performing ones usually leverage on random-forests or CNNs to predict match reliability. Therefore, a proper amount of labeled data is required to effectively train such confidence measures. Being such ground-truth labels not always available in practical applications, in this paper we propose a methodology suited for training confidence measures in a self-supervised manner. Leveraging on a pool of properly selected conventional measures, we automatically detect a subset of very reliable pixels as well as a subset of erroneous samples from the output of a stereo algorithm. This strategy provides labels for training confidence measures based on machine-learning technique without ground-truth labels. Compared to state-of-the-art, our method is neither constrained to image sequences nor to image content. Experimental results on three challenging datasets with three stereo algorithms and three state-of-the-art confidence measures based on machine-learning techniques confirm the effectiveness of our proposal for self-supervised training.

1 Introduction

Accurate and dense depth estimation is a crucial step for several computer vision applications. Passive stereo, compared to active technologies such as LIDAR, *structured light* and *time-of-flight* sensors, is often the preferred choice thanks to its accuracy, low cost and its fitting with indoor and outdoor environments. Although most state-of-the-art stereo algorithms rely on the popular disparity optimization proposed by Semi Global Matching (SGM) [9], a recent trend to improve stereo accuracy, especially when facing with challenging environments [6, 17, 28], consists in exploiting the additional information provided by *confidence measures* [20, 24, 29, 31] encoding the degree of uncertainty of depth data. The common ground in these works consists in inferring a confidence measure trained on labeled data and

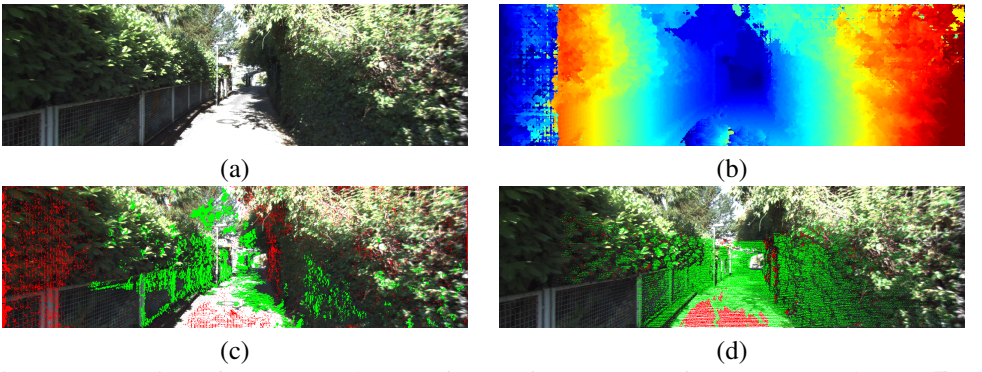

Figure 1: Overview of our proposal. (a) Reference frame 000122 from KITTI 12 dataset [6], (b) disparity map generated by SGM algorithm [9], (c) labels inferred by our method and (d) labels assigned comparing the output of SGM with ground-truth data. In green and red, respectively, *correct* and *wrong* labels.

its deployment to adjust cost volume (i.e., the pool of matching costs) for improved stereo accuracy. Focusing on confidence estimation, machine-learning approaches proved to be more effective than traditional ones, reviewed and evaluated in [10] and more recently in [22], as reported in [8, 20, 24, 31] for methods based on random-forest and in [21, 25, 29] for those leveraging on Convolutional Neural Networks (CNNs). Despite their differences, all these methods predict match reliability and thus, both conventional and machine-learning approaches, will be referred to as confidence measure.

Regardless of their specific deployment purpose, confidence estimation techniques based on machine-learning require a significant amount of training samples obtained from *ground-truth* data. In general, the higher amount and variety of labeled data available, the more effective the confidence estimation is. However, excluding a tedious and time consuming manual labeling, accurate ground-truth labels require either not trivial setup based on structured light, as described in [28], or expensive and appropriately registered active sensors, typically LIDAR, as done in [6, 17]. The first strategy provides dense (i.e., available for each point) ground-truth labels but it is only suited for still scenes acquired in indoor environments while the latter one enables to determine sparse ground-truth data from any indoor and outdoor environment. To overcome these issues, synthetic datasets have been recently deployed to train end-to-end stereo methods based on CNNs [16] with satisfactory results. However, such method requires an additional fine tuning on large labeled real data (*e.g.*, the whole 194 images of the KITTI 2012 training dataset in [16]) to achieve top performance on standard datasets. Thus, regardless of the desired goal, self-supervised and accurate labeling of disparity is crucial when dealing with machine-learning algorithms that require, for a specific application domain, a large amount of training samples as would occur in most practical circumstances.

To this aim Mostegel et al. [19] proposed an automatic technique, referred to as SELF, capable to automatically assign labels to train confidence measures by leveraging on contradictions and consistencies between disparity maps generated by the same stereo algorithm from multiple view points. This self-supervised approach proved to be very effective but it intrinsically suffers of two strong limitations. Firstly, it requires image sequences which are not always available. For instance, the Middlebury 2014 dataset [28] does not provide

such data at all. Moreover, this method accounts for camera ego-motion but it does not enable to detect labels belonging to moving subjects, such as cars or pedestrians, in the sensed environment.

Therefore, to overcome these issues we propose an approach to automatically generate, in a self-supervised manner, labels for training confidence measures without any of the aforementioned constraints. Our method, given a disparity map generated by a stereo algorithm, assigns a *correct* label to highly confident points and a *wrong* label to poorly reliable disparity measurements leveraging on the joint estimation provided by a pool of conventional confidence measures which do not require any training phase. Figure 1 summarizes our proposal. Given a stereo pair (a) and the disparity map generated by a stereo algorithms (b), we determine (c) training labels by assigning correct (green) or wrong (red) labels according to the joint confidence estimation carried out by means of conventional measures. In this very image, compared to ground-truth, our method correctly estimates 97.57% of correct and wrong labels. In the same figure, (d) shows for the same disparity map the intersection with ground-truth points.

We assess the performance of our self-supervised labeling approach on three challenging datasets (KITTI 12 [6], KITTI 15 [17] and Middlebury 2014 [28], referred to as MIDD 14) with three stereo algorithms characterized by different accuracy (block-matching, MC-CNN [32] and SGM [9]) by training on labels inferred by our method three state-of-the-art confidence measures [24, 25, 29] based on machine-learning. Our experimental evaluation with three state-of-the-art confidence measures clearly highlights that, using the same images for training, the proposed method not only provides an unconstrained labeling strategy with respect to SELF [19] but it also yields much more accurate confidence estimation.

2 Related work

The literature concerning confidence measures is relevant to our work and this topic has been reviewed and evaluated in [3, 4, 10]. In particular, [10] categorizes conventional confidence approaches according to the input cue processed and quantitatively evaluates their performance on two datasets by exploiting ROC curve analysis. More recent works in this field, reviewed and evauated in [23], deploy machine-learning techniques. In [8, 20, 24, 31] by feeding a random-forest with hand-crafted features and in [25, 29] by analyzing with a CNN the raw disparity map [25], features extracted from it [29] or the entire cost volume [30]. Currently, [21, 24, 25, 29] are top-performing approaches for confidence estimation and in [27] was shown how to improve the accuracy of a confidence prediction by exploiting its local consistency with a CNN. Concerning confidence measures for embedded systems, in [23] were reviewed and evaluated approaches compatible with constrained architectures. Recent works also proved the potential of confidence measures to achieve better results from stereo. Specifically, in [31] as input cue for disparity inference based on MRF while in other cases to improve the effectiveness of the SGM algorithm modulating its raw matching costs [20], weighting the contribution of multiple scanline optimizations [24], dynamically adjusting P1 and P2 parameters [29] or for disparity refinement [7]. Finally, in [15, 18], confidence measures have been deployed for sensor fusion.

Concerning stereo, MC-CNN [32] represents the first successfully attempt to deploy deep learning for disparity inference by training a CNN to predict raw matching costs refined by a conventional processing pipeline. Other approaches following this strategy are [1, 14] while in [26] a CNN was trained to combine the outcome of multiple stereo algorithms. Mayer et

al. [16] proposed the first end-to-end approach for disparity prediction by training a CNN on a large synthetic dataset and fine-tuning their method on realistic dataset with ground-truth. In this field, current state-of-the-art is represented by Kendall et al. [10]. It deploys a very deep architecture trained end-to-end: extracting features to build a cost volume, processing such data by means of 3D convolutions and finally adopting a learned WTA strategy.

A common issue with machine-learning techniques for disparity and confidence estimation is the requirement of training data. In fact, most methods need a large amount of ground-truth labeled samples to achieve the best performance. Thus, self-supervised learning earn increasing importance to overcome this issue when large datasets are not available. Garg et al. [6] proposed a self-supervised learning framework applied to a CNN deployed for single-camera depth estimation. Long et al. [13] proved how to learn image matching by training a CNN on video sequences. Given two images, the network is trained to infer a third one, then gradients with respect to input frames are computed and their response is used to find corresponding pixels. Finally, Mostegel et al. [19] proposed a methodology to automatically generate training data from stereo sequences, reasoning on contradictions and consistencies between disparity maps obtained from different view points and testing their strategy to train machine-learning based confidence measures. To the best of our knowledge, this is the only method to obtain for this task training labels in a self-supervised manner and thus it represents the most relevant approach concerned with our work.

3 Self-supervised labeling

In this section we outline our proposal to automatically determine training labels from stereo pairs in order to obtain a distribution of training labels as much as possible similar to GT data. The fundamental underlying assumption made by our method concerns the capability of a combination of *hand-crafted* confidence measures to discriminate between correct and wrong disparity assignments generated by a stereo algorithm. This selection procedure allows us to obtain two distinct labels, *correct* and *wrong*, that can be used as training samples for state-of-the-art confidence measures based on machine-learning. The primary goal of this method is to find a set of values as accurate as possible with the aim of reducing the number of false positive and false negative labels which could negatively affect training and consequently inference. Since we want to avoid a *chicken-and-egg* situation we can't rely on machine-learning confidence estimation for label selection and thus a careful choice of traditional confidence measures is mandatory.

The effectiveness of a specific confidence measure is quantitatively assessed by means of a ROC curve analysis [10, 22] according to a standard procedure in this field [8, 20, 21, 24, 25, 27, 29, 31]. This strategy enables to determine how well a confidence estimator can discriminate between correct and wrong matches. The behavior of the curve itself encodes several important aspects of a confidence measure. For example, a flat portion of the curve indicates a large amount of pixels sharing the same estimated confidence. The extensive evaluation reported in [10] showed how different measures behave differently according to the processed cues as well as the adopted strategy. In particular, for the same pixel, different measures typically provide contradictory scores. This fact has been successfully exploited to infer much more effective confidence measures analyzing with random-forest [8, 20, 24, 31] or a CNN [21] a pool of not very effective confidence measures.

Our strategy relies on a set of conventional, yet according to the literature [2, 10, 12, 20, 22, 31] reliable, confidence measures to automatically generate classification labels with a

distribution as much as possible similar to GT data required to train state-of-the-art measures based on machine-learning. Differently from [19], our proposal does not enforce any constraint on the input data being it suited for image sequences, for uncorrelated stereo pairs as well as for scenes containing moving objects.

3.1 Confidence measures for label selection

In this section we review the confidence measures adopted by our method. We carefully selected them according to the voting technique deployed to generate labels, explained in detail in section 3.2. Given the cost curve provided by a stereo algorithm for a pixel $\mathbf{p}(x,y)$, the chosen confidence measures process (a subset of) cues such as the minimum cost $c_1(\mathbf{p}) \equiv c_1(\mathbf{p}, d_1(\mathbf{p}))$ at disparity hypothesis $d_1(\mathbf{p})$, the second smallest local minimum as $c_{2m}(\mathbf{p}) \equiv c_{2m}(\mathbf{p}, d_{2m}(\mathbf{p}))$ at disparity hypothesis d_{2m} (and, in general, the cost for a certain disparity hypothesis d as $c_d(\mathbf{p})$), the disparity value $\mathcal{D}(\mathbf{p})$ assigned by *winner-takes-all* strategy to $\mathbf{p}$ and its corresponding pixel on the right image referred to as $\mathbf{p}'$, having disparity $\mathcal{D}^R(\mathbf{p}')$. We denote as $N_\mathbf{p}$ a squared patch centered on pixel $\mathbf{p}$ (of size 25×25 in our experiments).

- **Average Peak Ratio (APKR)** [12]: computed by processing the ratio between $c(\mathbf{q}, d_{2m}(\mathbf{p}))$ and $c(\mathbf{q}, d_1(\mathbf{p}))$, averaged on a squared neighborhood.

$$APKR(\mathbf{p}) = \frac{1}{|N_\mathbf{p}|} \sum_{\mathbf{q} \in N_\mathbf{p}} \frac{c(\mathbf{q}, d_{2m}(\mathbf{p}))}{c(\mathbf{q}, d_1(\mathbf{p}))} \tag{1}$$

- **Left-Right Consistency (LRC)** [10, 22]: obtained by comparing the disparity of pixel $\mathbf{p}$ with the corresponding point $\mathbf{p}'$ on right disparity map.

$$LRC(\mathbf{p}) = \begin{cases} 0, & \text{if } \mathcal{D}(\mathbf{p}) \neq \mathcal{D}^R(\mathbf{p}') \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

- **Median deviation of disparity (MED)** [31]: represents the difference between disparity D on pixel $\mathbf{p}$ and the median disparity computed on a squared neighborhood:

$$MED(\mathbf{p}) = \begin{cases} 0, & \text{if } \mathcal{D}(\mathbf{p}) \neq median_{N_\mathbf{p}}(\mathcal{D}(\mathbf{p})) \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

- **Uniqueness Constraint (UC)** [2]: a binary measure that encodes with low confidence points colliding on the same pixel $\mathbf{p}'$ in the right image thus violating the *uniqueness* constraint:

$$UC(\mathbf{p}) = \begin{cases} 0, & \text{if } \mathbf{p} \in Q \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

  being Q the set of pixels matching the same point on the right image.

- **Winner Margin (WMN)** [10, 22]: obtained by processing the difference between local minimum c_{2m} and minimum cost c_1, normalized by the sum of costs over the entire disparity range.

$$WMN(\mathbf{p}) = \frac{c_{2m}(\mathbf{p}) - c_1(\mathbf{p})}{\sum_d c_d(\mathbf{p})} \tag{5}$$

- **Distance to Left Border** (**DLB**) [20]: distance from the left border of the image, thresholded to the maximum disparity value $\mathcal{D}_{max}$ set for the stereo algorithm:

$$DLB(\mathbf{p}) = \begin{cases} 0, & \text{if } x < \mathcal{D}_{max} \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

3.2 Label selection strategy

Given a disparity map D generated by a stereo algorithm, we want to reliably assign on subset of points labels $\mathcal{L} = \{L_0, L_1\}$ standing, respectively, for wrong and correct. From each of the confidence measures previously described, we obtain a map $\mathcal{C}$ assigning values $\in [0,1]$ to each point $\in D$. We define two sets of points $\mathcal{C}_0$ and $\mathcal{C}_1$ one for each label L_0 and L_1. For binary confidence measures we simply assume as correct points $\mathbf{p}$ with $\mathcal{C}(\mathbf{p}) = 1$ and as wrong those with $\mathcal{C}(\mathbf{p}) = 0$ while for the others the choice is made by sorting all points $\in D$ in ascending order of confidence and then defining the two sets as:

$$\mathcal{C}_0 = \{\mathbf{p} \in D | 0 \leq \mathcal{C}(\mathbf{p}) \leq \delta_0\}, \quad \mathcal{C}_1 = \{\mathbf{p} \in D | 1 - \delta_1 \leq \mathcal{C}(\mathbf{p}) \leq 1\} \tag{7}$$

with (δ_0, δ_1) representing portions of the entire disparity map, corresponding to the least ($\mathcal{C}_0$) and most ($\mathcal{C}_1$) confident pixels. For example, with $(\delta_0, \delta_1) = (0.2, 0.2)$, $\mathcal{C}_0$ will group the 20% pixels having lowest confidence value and $\mathcal{C}_1$ the 20% having highest scores.

By following this strategy for each $\mathcal{C}$ in a pool $\mathcal{P} = \{\mathcal{C}', \mathcal{C}'', ..\}$ of confidence measures, we obtain two ensembles $\mathcal{P}_0 = \{\mathcal{C}'_0, \mathcal{C}''_0, ...\}$ and $\mathcal{P}_1 = \{\mathcal{C}'_1, \mathcal{C}''_1, ...\}$ for the two labels L_0 and L_1. We combine the different labeling hypothesis $\in \mathcal{P}$ provided by the measures to obtain the final sets $\mathcal{G}_0$, $\mathcal{G}_1$ as follows:

$$\mathcal{G}_0 = \bigcap_{\mathcal{C}^k \in \mathcal{P}_0} \mathcal{C}^k_0, \qquad \mathcal{G}_1 = \bigcap_{\mathcal{C}^k \in \mathcal{P}_1} \mathcal{C}^k_1 \tag{8}$$

According to this strategy, in order to reduce false positives and negatives originated by each single measure, only pixels classified by all the confidence measures as either correct or wrong are used for labeling. On the other hand, this conservative strategy also reduces the amount of pixels for which our method provides labels. Our conservative selection strategy aims at obtaining very accurate labels comparable to those provided by GT data.

4 Experimental Results

In this section, we assess[1] the effectiveness of our proposal with three datasets and three stereo algorithms by training three state-of-the-art confidence measures with the labels generated by our method, the ones generated by SELF [19] as well as using ground-truth data and comparing their performance by means of ROC analysis. Regarding the datasets, we consider KITTI 12 [6], KITTI 15 [17] and MIDD 14 [28]. As confidence measures we choose the three top-performing methods known in literature: O1 [24], CCNN [25] and PBCP [29]. The choice of these measures was driven by their effectiveness with respect to all other machine learning approaches. In particular, all of them proved to outperform the work of [20]. Concerning the stereo algorithms, we consider three approaches characterized

[1]For SELF [19], O1 [24], CCNN [25] and MC-CNN [33] we used the code available in the authors' web site while for PBCP [29], CENSUS and SGM we implemented them following the description available in each paper.

KITTI 12	CENSUS		MC-CNN		SGM	
Method	A	D / D∩GT	A	D / D∩GT	A	D / D∩GT
SELF [19]	88.9%	33.8% / 38.0%	85.4%	29.4% / 30.7%	81.3%	21.5% / 23.2%
Prop.	98.5%	8.4% / 12.5%	97.0%	12.4% / 13.3%	88.6%	12.5% / 14.6%

Table 1: Analysis of training labels inferred on 8 sequences of KITTI 12. For SELF [19] and our proposal we report the accuracy A for the predicted labels (computed for points with available ground-truth), the average density D on the 8 sequences, the intersection between the density of labels inferred by the two methods and the 8 images with ground-truth (D∩GT). The average density of KITTI 12 ground-truth data on the 8 images is 19.5%.

by different performance. The popular, yet not very effective, block matching algorithm, referred to as CENSUS, aggregating costs (computed by means of Hamming distance on census transformed images) with a 5×5 box-filter. As representative of algorithms with high accuracy we use MC-CNN [32], considering the matching costs computed on patches (9×9 on KITTI 12 and KITTI 15 and 11×11 on MIDD 14 and using the weights provided by the authors), and SGM [9] in a eight scanlines implementation using for data term the same CENSUS aggregated costs and for parameters P1 and P2, respectively, 0.03 and 3 (being matching costs normalized).

4.1 Evaluation protocol and training data

In this field, ROC curves [10, 22] are commonly deployed to assess how reliable is a confidence measure by sorting pixels $\in \mathcal{D}$ according to their scores and computing error rates on subsets sampled with increasing size. If sorted in descending order, an effective confidence measure should sample correct pixels first and, then, outliers. Thus, the Area Under the Curve (AUC) quantitatively summarizes the behavior on the entire disparity map, enabling to compare different measures. The lower is the area, the more effective is the prediction. The optimal AUC is obtained as $\varepsilon + (1-\varepsilon)\ln(1-\varepsilon)$, being ε the percentage of outliers in the disparity map $\mathcal{D}$.

Confidence measures are trained in most works in this field [8, 20, 24, 25, 31] by selecting eight stereo pairs from KITTI 12 dataset: 43, 71, 82, 87, 94, 120, 122 and 180^{th}. These images with ground-truth labels provide about 724K training samples. According to SELF [19], on the extended eight sequences available on KITTI 12 corresponding to the 8 stereo pairs 43, 71, 82, 87, 94, 120, 122 and 180^{th}, we generate training labels following the protocol described by the authors. For all considered sequences there are available 21 stereo pairs, excluding 82^{th} containing only 16. On such 163 stereo pairs SELF extracts a huge amount of training labels: about 25M for CENSUS, 22M for MC-CNN and 16M for SGM. For a fair comparison, we generate labels with our method from the same sequences. However, differently from SELF, we point out that our method is not constrained to sequences but we use for the aforementioned reason the same input data to generate our training labels. In fact, taking the same number of stereo pairs from different scenes would favour our approach making the comparison unfair. Overall, our framework provides from the eight sequences about 6M training labels for CENSUS and 9M for MC-CNN and SGM.

Despite the significantly lower amount of labels generated by our proposal with respect to SELF, observing Table 1 we can notice that our training samples are always more accurate. This fact highlights that our proposal significantly reduces the percentage of wrong assignments to $\mathcal{G}_0$ and $\mathcal{G}_1$ trading accuracy for density. Moreover, it is worth to note that

KITTI 12	CENSUS (ε=38.6%)			MC-CNN (ε=16.9%)			SGM (ε=9.1%)		
measure	GT	[19]	Prop.	GT	[19]	Prop.	GT	[19]	Prop.
O1 [24]	0.116	0.165	0.163	0.025	0.046	0.042	0.016	0.031	0.022
CCNN [25]	0.118	0.250	0.128	0.028	0.089	0.029	0.032	0.084	0.023
PBCP [29]	0.125	0.201	0.138	0.029	0.044	0.040	0.029	0.037	0.035
APKR [12]		0.166			0.048			0.030	
opt.		0.094			0.017			0.005	
KITTI 15	CENSUS (ε=35.4%)			MC-CNN (ε=15.4%)			SGM (ε=13.7%)		
measure	GT	[19]	Prop.	GT	[19]	Prop.	GT	[19]	Prop.
O1 [24]	0.109	0.172	0.147	0.031	0.059	0.046	0.021	0.038	0.027
CCNN [25]	0.113	0.266	0.120	0.036	0.102	0.035	0.044	0.072	0.029
PBCP [29]	0.122	0.209	0.151	0.035	0.053	0.047	0.031	0.035	0.037
APKR [12]		0.147			0.049			0.036	
opt.		0.083			0.019			0.007	
MIDD 14	CENSUS(ε=37.8%)			MC-CNN (ε=26.7%)			SGM (ε=26.9%)		
measure	GT	[19]	Prop.	GT	[19]	Prop.	GT	[19]	Prop.
O1 [24]	0.126	0.180	0.154	0.073	0.125	0.097	0.085	0.133	0.102
CCNN [25]	0.128	0.254	0.123	0.072	0.179	0.069	0.122	0.216	0.088
PBCP [29]	0.119	0.169	0.123	0.067	0.084	0.078	0.145	0.148	0.148
APKR [12]		0.137			0.074			0.100	
opt.		0.090			0.046			0.045	

Table 2: Average AUCs on the 3 datasets (from top to bottom: KITTI 12, KITTI 15 and MIDD 14). Evaluation of the 3 confidence measures with 3 algorithms (CENSUS, MC-CNN, SGM), trained on ground-truth data (GT), on labels obtained by SELF [19] and by our proposal. We also include in the table a single AUC for each algorithm concerned with APKR not affected at all by training labels. We also report the average error ε on each dataset computed with error bound set to 3, for KITTI datasets, and set to 1 for MIDD 14.

KITTI 12 provides, on the 8 images, ground-truth labels only for 19.5% of points. On the 8 sequences SELF always generates a larger percentage of labels, parameter D in the table, compared to our method. We can also notice from D∩GT that our method selects a larger percentage of points not overlapping with ground-truth data with respect to SELF. This fact potentially allows us to include more points in regions not covered by LIDAR as shown in Figure 1 in the left and upper side of the disparity map. Moreover as reported in Figure 2, we observed that with respect to our proposal SELF provides a limited amount of correct samples for farther points in the disparity map. All these facts might explain the overall best performance of our strategy and why, in some circumstances, it allows us to achieve more accurate results with respect to deploy ground-truth labels for training confidence measures as will be detailed in the next section.

4.2 Quantitative evaluation and analysis of training data

In this section we exhaustively compare our proposal with SELF [19] on three datasets KITTI 12, KITTI 15 and MIDD 14 and three algorithms for training the three state-of-the-art confidence measures O1 [24], CCNN [25] and PBCP [29] trained on labels inferred from eight sequences belonging to KITTI 12.

Moreover, we compare the performance of the same confidence measures trained on

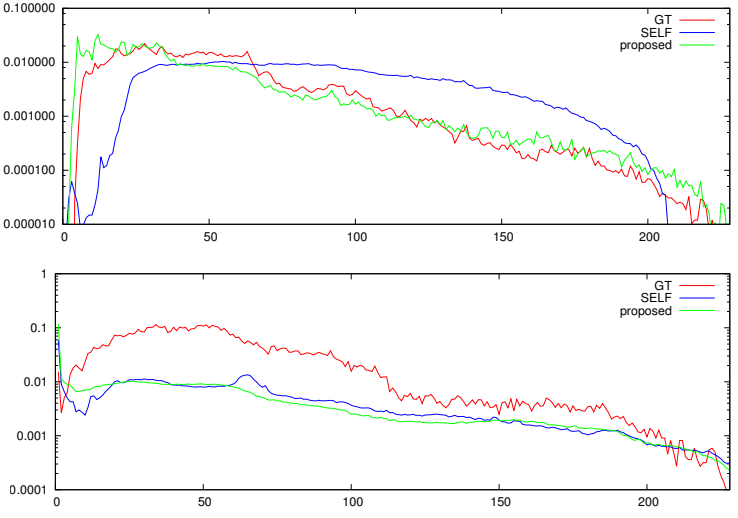

Figure 2: Distribution of training labels with SGM for SELF [19] in blue and the proposed method in green. In red the distribution of GT labels, independent of the stereo algorithm. (Top) Distribution of *correct* and (bottom) *wrong* labels within the disparity range.

labels extracted from the corresponding eight stereo pairs with ground-truth data available in KITTI 12. Detailed experimental results are reported in Table 2. We include in our evaluation APKR [12], the most effective confidence measure within the pool of confidence measures deployed for selecting labels as described in Section 3.1. Being such method independent of the training labels we report in the table a single AUC for APKR. On KITTI 12, our proposal always enables more effective training of confidence measures with respect to SELF. In particular, with CCNN and in most cases with PBCP, our method performs much better. Confidence measures trained with our method are more reliable than APKR in 8 out of 9 times while SELF yields better results only in 3 out of 9 times. Compared to training confidence measures on GT labels, SELF is always less reliable while our proposal with SGM and CCNN yields significantly better results. It is worth to note that, although the accuracy of our labels is higher compared to SELF, the amount of samples provided by our method for training is much lower.

The cross validation on KITTI 15 shows that our method is always more effective than SELF. Similarly to the results reported for KITTI 12, the validation on KITTI 15 highlights that CCNN has better performance when trained with our labels with respect to train on SELF. This trend is also confirmed with PBCP in many cases. APKR achieves better AUCs compared to our method in 2 out of 9 cases while SELF in 8 out of 9 cases. Compared to training on GT labels, our proposal enables to achieve better results in two cases (with CCNN) while SELF never yields better confidence estimation. The cross validation on MIDD 14 highlights, once more, that our self-labeling approach outperforms SELF excluding the test with CCNN trained on labels generated with SGM where the two methods have equivalent performance very similar to the AUC obtained training on GT labels. Compared to APKR, our method is better in 4 out of 9 situations (with any stereo algorithm training CCNN and, with CENSUS, training PBCP) while SELF is always outperformed by this method. Moreover, we point out that CCNN trained with our proposal yields always

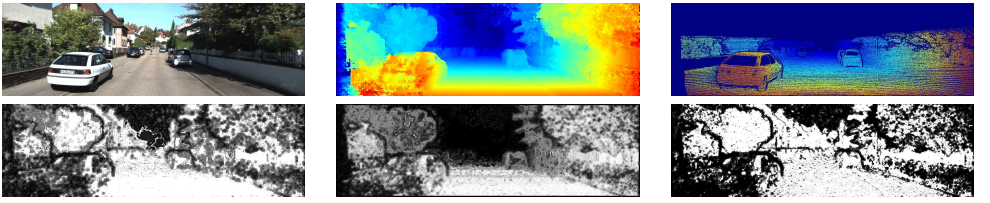

Figure 3: Qualitative results with the SGM algorithm on frame 000006 belonging to KITTI 12. At the top, from left to right, reference image, disparity computed by SGM and GT. At the bottom, we report for O1 [24] confidence maps obtained, from left to right, training on GT data, SELF [19] and the proposed method.

to more accurate results with respect to training on GT labels while this fact never holds for SELF. The experimental results reported in Table 2 confirm that our proposal enables more effective training of confidence measures with respect to SELF as well as to a better generalization to new data. Moreover, training on labels generated by our method allows us, in most cases, to obtain confidence measures (in particular with those based on CNNs, CCNN and PBCP) with performance comparable, and sometimes even better, than training the same measures on ground-truth labels. In Figure 2 we compare the distribution of *correct* and *wrong* training labels obtained by SELF and our proposal with KITTI 12. We also report the distribution of GT data. Observing the figures we can observe that our method generates training labels more similar to GT data. Moreover, we can notice how SELF provides very few positive labels for higher and lower disparity values especially dealing with correct labels. Figure 3 shows qualitative results for O1 confidence measure and SGM algorithm, obtained by training the measure on data from GT, SELF and our method. Finally, excluding disparity and confidence computation, on a i7 CPU, with our method we automatically extracted the training samples from 163 images of KITTI 12 in 76 seconds.

5 Conclusions

In this paper we have proposed a novel self-supervised strategy to train confidence measures based on machine-learning. Compared to state-of-the-art methods our proposal is more general and neither constrained to image sequences nor to scene content. It generates training labels by leveraging on a pool of appropriately combined conventional confidence measures. The experimental results reported confirm that our strategy improves state-of-the-art by selecting more accurate labels thus enabling better confidence estimation when training confidence measures based on machine-learning on self-generated data. Moreover, in particular with CNN-based confidence measures, it also provides competitive results with respect to ground-truth. This fact confirms our method can be deployed to train confidence measures from unlabeled stereo pairs, a circumstance frequently occurring in practical applications. Future work is aimed at further improving the proposed labeling selection strategy.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] Zhuoyuan Chen, Xun Sun, Liang Wang, Yinan Yu, and Chang Huang. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 972–980, 2015.

[2] Luigi Di Stefano, Massimiliano Marchionni, and Stefano Mattoccia. A fast area-based stereo matching algorithm. *Image and vision computing*, 22(12):983–1005, 2004.

[3] Geoffrey Egnal and Richard P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 24(8):1127–1133, 2002.

[4] Geoffrey Egnal, Max Mintz, and Richard P. Wildes. A stereo confidence metric using single view imagery. In *PROC. VISION INTERFACE*, pages 162–170, 2002.

[5] Ravi Garg, B. G. Vijay Kumar, Gustavo Carneiro, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 740–756, 2016.

[6] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, sep 2013.

[7] Spyros Gidaris and Nikos Komodakis. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR. Proceedings*, pages 305–312, 2013.

[9] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2): 328–341, feb 2008.

[10] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 2121–2133, 2012.

[11] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression, 2017.

[12] S. Kim, D. g. Yoo, and Y. H. Kim. Stereo confidence metrics using the costs of surrounding pixels. In *2014 19th International Conference on Digital Signal Processing*, pages 98–103, Aug 2014.

[13] Gucan Long, Laurent Kneip, Jose M. Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 434–450, 2016.

[14] W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. In *Proc. CVPR*, 2016.

[15] Giulio Marin, Pietro Zanuttigh, and Stefano Mattoccia. Reliable fusion of tof and stereo depth driven by confidence measures. In *ECCV 2016*, pages 386–401, 2016.

[16] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] Paul Merrell, Amir Akbarzadeh, Liang Wang, Jan Frahm, and Ruigang Yang David Nistér. Real-time visibility-based fusion of depth maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[19] Christian Mostegel, Markus Rumpler, Friedrich Fraundorfer, and Horst Bischof. Using self-contradiction to learn confidence measures in stereo vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Min-Gyu Park and Kuk-Jin Yoon. Leveraging stereo matching with learning-based confidence measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] M. Poggi, F. Tosi, and S. Mattoccia. Even more confident predictions with deep machine-learning. In *12th IEEE Embedded Vision Workshop, CVPR 2017 workshop*, Jul 2017.

[22] 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)*, Oct 2017.

[23] M. Poggi, F. Tosi, and S. Mattoccia. Efficient confidence measures for embedded stereo. In *19 International Conference on Image Analysis and Processing (ICIAP 2017)*, Sep 2017.

[24] Matteo Poggi and Stefano 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.

[25] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *Proceedings of the 27th British Conference on Machine Vision, BMVC*, 2016.

[26] Matteo Poggi and Stefano Mattoccia. Deep stereo fusion: combining multiple disparity hypotheses with deep-learning. In *Proceedings of the 4th International Conference on 3D Vision, 3DV*, 2016.

[27] Matteo Poggi and Stefano Mattoccia. Learning to predict stereo reliability enforcing local consistency of confidence maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nesic, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition: 36th German Conference, GCPR 2014*, pages 31–42.

[29] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *British Machine Vision Conference (BMVC)*, 2016.

[30] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1621–1628. IEEE, 2014.

[32] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016.