Quantitative evaluation of confidence measures in a machine learning world

Matteo Poggi, Fabio Tosi, Stefano Mattoccia
University of Bologna
Department of Computer Science and Engineering (DISI)
Viale del Risorgimento 2, Bologna, Italy
{matteo.poggi8, fabio.tosi5, stefano.mattoccia}@unibo.it

Abstract

Confidence measures aim at detecting unreliable depth measurements and play an important role for many purposes and in particular, as recently shown, to improve stereo accuracy. This topic has been thoroughly investigated by Hu and Mordohai in 2010 (and 2012) considering 17 confidence measures and two local algorithms on the two datasets available at that time. However, since then major breakthroughs happened in this field: the availability of much larger and challenging datasets, novel and more effective stereo algorithms including ones based on deep learning and confidence measures leveraging on machine learning techniques. Therefore, this paper aims at providing an exhaustive and updated review and quantitative evaluation of 52 (actually, 76 considering variants) state-of-the-art confidence measures - focusing on recent ones mostly based on random-forests and deep learning - with three algorithms on the challenging datasets available today. Moreover we deal with problems inherently induced by learning-based confidence measures. How are these methods able to generalize to new data? How a specific training improves their effectiveness? How more effective confidence measures can actually improve the overall stereo accuracy?

1. Introduction

Although depth from stereo still represents an open problem [5, 23, 37], in recent years this field has seen notable improvements concerning the effectiveness of such algorithms (*e.g.*, [47, 40]) and confidence measures, aimed at detecting unreliable disparity assignments, proved to be very effective cues when plugged in stereo vision pipelines as shown in [41, 28, 31, 40]. However, shortcomings of stereo algorithms have been emphasized by the availability of very challenging datasets with ground-truth such as KITTI 2012 (K12) [5], KITTI 2015 (K15) [23] and Middlebury 2014 (M14) [37]. Thus, the ability to reliably predict failures of a stereo algorithm by means of a confidence measure is fundamental and many approaches have been proposed for this purpose. Hu and Mordohai [13] exhaustively reviewed confidence measures available at that time, with two variants of a standard local algorithm, and defined a very effective metric to evaluate their effectiveness on the small and mostly unrealistic dataset [39] with ground-truth available. However, since then there have been major breakthroughs in this field:

- Novel and more reliable confidence prediction methods, in particular those based on random-forests [8, 41, 28, 31] and deep learning [32, 40]
- Much larger datasets with ground-truth depicting very challenging and realistic scenes acquired in indoor [37] and outdoor environments [5, 23]
- Novel and more effective stereo algorithms, some leveraging on deep learning techniques [47, 22], more and more often coupled with confidence measures [40, 31, 28]. Moreover, in recent years, SGM [9] became the preferred disparity optimization method for most state-of-the-art stereo algorithms (*e.g.*, [47, 40])

Considering these facts, we believe that this field deserves a further and deeper analysis. Therefore, in this paper we aim at i) extending and updating the taxonomy provided in [13] including novel confidence measure and in particular those based on machine learning techniques, ii) exhaustively assessing their performance on the larger and much more challenging datasets [23, 37] available today, iii) understanding the impact of training data on the effectiveness of confidence measures based on machine learning, iv) assessing their performance when dealing with new data and state-of-the-art stereo algorithms, v) and evaluating their behavior when plugged into a state-of-the-art stereo pipeline.

Although our focus is mostly on approaches based on machine learning, for completeness, we include in our taxonomy and evaluation any available confidence measure.

Overall, we assess the performance of 52 measures, actually 76 considering their variants, providing an exhaustive evaluation of state-of-the-art in this field with three stereo algorithms on the three challenging datasets with ground-truth K12, K15 and M14 available today.

2. Related work

The most recent taxonomy and evaluation of confidence measures for stereo was proposed by Hu and Mordohai [13]. They exhaustively categorized the 17 confidence measures available at that time in six categories according to the cues exploited to infer depth reliability. Moreover, they proposed an effective metric to clearly assess the effectiveness of confidence prediction based on area under the curve (AUC) analysis and quantitatively evaluated the considered measures on former indoor Middlebury [39, 12] and outdoor *Fountain P11* [42] datasets with a standard local algorithm using *sum of absolute differences* (SAD) and *normalized cross correlation* (NCC) as matching costs.

However, since then novel confidence measures were proposed [14, 15, 25, 2, 7, 8, 41, 28, 31] and more importantly this field was affected by methodologies inspired by machine learning. In their seminal work, Hausler et al. [8] proposed to infer match reliability by feeding a random forest, trained for classification, with multiple confidence measures showing that this fusion strategy yields much better performance with respect to any other considered confidence measure. Following this strategy, the reliability of confidence measures was further improved in [41] and [28] considering more effective features. In this context, [31] enables to infer a confidence measure leveraging only features extracted in constant time from the left disparity map. Differently from [8], in [41, 28, 31] the random-forests are trained in regression mode. Concerning methodologies based on Convolutional Neural Networks (CNN), Seki and Pollefeys [40] proposed to infer a confidence measure by processing features extracted from the left and right disparity maps while Poggi and Mattoccia [32] learned from scratch a confidence measure by feeding to a CNN the left disparity map. Moreover, in [35] was proposed a method to combine multiple hand-crafted cues and in [33] a strategy to improve confidence accuracy exploiting local consistency. Concerning unsupervised training of confidence measures, Mostegel et al. [27] proposed to determine training labels exploiting contradictions between multiple depth maps computed from different viewpoints while Tosi et al. [43] leveraging on a pool of confidence measures.

This field has also seen the deployment of confidence measures plugged into stereo vision pipelines to improve the overall accuracy as proposed in [41, 28, 31, 40, 29, 36, 16, 26, 6], to deal with occlusions [9, 25] or to improve accuracy near depth discontinuities [4]. Most of these approaches are aimed at improving the accuracy of Semi Global Matching (SGM) [9] algorithm exploiting as cue an estimated match reliability. Confidence measures have been effectively deployed for sensor fusion combining depth maps from multiple sensors [20, 24]. Finally, confidence measures suited for embedded stereo systems have been analyzed in [34].

Recent years have also witnessed the availability of very challenging datasets depicting indoor, such as the M14 [37], and outdoor environments, such as K12 [5] and K15 [23]. Differently from former standard dataset [39] used to test algorithms, the novel ones clearly emphasize that stereo is still an open research problem. This fact also paved the way to most recent trend in stereo vision aimed at tackling stereo with CNNs. In this context [47] Zbontar and Le Cun proposed the first successful attempt to infer an effective matching cost from a stereo pair with a CNN now deployed by almost any top-performing stereo method on K12, K15 and M14 datasets. Following this strategy Chen et al. [1] and Luo et al. [18] proposed very efficient architectures enabling real-time stereo matching while [30] enables to combine multiple disparity maps with a CNN. A further step forward, aimed at departing from a conventional stereo pipeline, is represented by Mayer et al. [22]. In this case, given a stereo pair, the left-right stereo correspondence is regressed from scratch with a CNN trained end-to-end.

3. Taxonomy of confidence measures

Despite the large number of confidence measures proposed, all of them process (a subset of) information concerning the cost curve, the relationship between left and right images or disparity maps. Following [13], confidence measures can be grouped into categories according to their input cues. To better clarify which cues are processed by each single measure we introduce the following notation. Given a stereo pair made of left (L) and right (R) images, a generic stereo algorithm assigns a cost curve c to each pixel of L. We denote the minimum of such curve as c_1 and its corresponding disparity hypothesis as d_1. We refer to the second minimum of the curve as c_2 (and to its disparity hypothesis as d_2), while c_{2m} denotes the second local minimum (it may coincide with c_2). In our taxonomy we group the considered 52 confidence measures (and their variants) in the following 8 categories.

3.1. Minimum cost and local properties of the cost curve

These methods analyze local properties of the cost curve encoded by c_1, c_2 and c_{2m}. As confidence values for each point, the *matching score measure* **(MSM)** [13] simply assumes the negation of minimum cost c_1. *Maximum margin* **(MM)** computes the difference between c_{2m} and c_1 while its variant *maximum margin naive* **(MMN)** [13] replaces

c_{2m} with c_2. *Non linear margin* **(NLM)** [7] computes a non linear transformation according to the difference between c_{2m} and c_1 while its variant *non linear margin naive* **(NLMN)** replaces c_{2m} with c_2. *Curvature* **(CUR) [13]** and *local curve* **LC** [44] analyze the behavior of the cost curve around the minimum c_1 and its two neighbors at (d_1-1) and (d_1+1) according two similar, yet different, strategies. *Peak ratio* **(PKR)** [10, 13] computes the ratio between c_{2m} and c_1. In one of its variants, *peak ratio naive* **(PKRN)** [13], c_{2m} is replaced with the second minimum c_2. In *average peak ratio* **(APKR)** [14] the confidence value is computed averaging PKR values on a patch. We include in our evaluation a further variant, based on the same patch-based average strategy adopted by APKR and referred to as *average peak ratio naive* **(APKRN)**. Similarly and respectively, *weighted peak ratio* **(WPKR)** [15] and *weighted peak ratio naive* **(WPKRN)**, average on a patch the original confidence measures PKR and PKRN with binary weights computed according to the reference image content. Finally, we include in this category two confidence measures belonging to the pool of features proposed in [8]. *Disparity ambiguity measure* **(DAM)** computes the distance between d_1 and d_2, while *semi-global energy* **(SGE)** relies on a strategy inspired by the SGM algorithm [9]. It sums, within a patch, the c_1 costs of points laying on multiple scanlines penalized, if their disparity is not the same of the point under examination, by P1 when the difference is 1 and by P2 (>P1) otherwise.

3.2. Analysis of the entire cost curve

Differently from previous confidence measures, those belonging to this category analyze for each point the overall distribution of matching costs. *Perturbation* **(PER)** [8] measures the deviation of the cost curve to an ideal one. *Maximum likelihood measure* **(MLM)** [21, 13] and *attainable likelihood measure* **(ALM)** [24, 13] infer from the matching costs a *probability density function* (pdf) with respect to an ideal c_1, respectively, equal to zero for MLM and to the actual c_1 for ALM. *Number of inflections* **(NOI)** [17] determines the number of local minima in the cost curve while *local minima in neighborhood* **(LMN) [14]** counts, on a patch, the number of points with local minimum at the same disparity d_1 of the examined point. *Winner margin measure* **(WMN)** [13] normalizes for each point the difference between c_{2m} and c_1 by the sum of all costs while its variant *winner margin measure naive* **(WMNN)** [13] adopts the same strategy replacing c_{2m} with c_2. Finally, *negative entropy measure* **(NEM)** [38, 13] relates the degree of uncertainty of each point to the negative entropy of its matching costs.

3.3. Left and right consistency

This category evaluates the consistency between corresponding points according to two different cues: one, symmetric, based on left and right maps and one, asymmetric, based only on the left map. Confidence measures adopting the first strategy are: *left-right consistency* **(LRC)** [3, 13], that assigns as confidence the negation of the absolute difference between the disparity of a point in L and its homologous point in R, and *left-right difference* **(LRD)** [13] that computes the difference between c_2 and c_1 divided by the absolute difference between c_1 and the minimum cost of the homologous point in R. We include in this category *zero-mean sum of absolute differences* **(ZSAD)** [8] that evaluates the dissimilarity between patches centered on homologous points in the stereo pair. It is worth pointing out that for LRC and ZSAD the full cost volume is not required. On the other hand, confidence measures based only on the analysis of the reference disparity map exploit the *uniqueness constraint*. *Asymmetric consistency check* **(ACC)** [25] and *uniqueness constraint* **(UC)** [2] detect the pool of multiple *colliding* points at the same coordinate in the right image. ACC verifies, according to a binary strategy, whether the candidate with the largest disparity in the pool has the smallest cost with respect to any other one while UC simply selects as valid the candidate with the minimum cost. Moreover, we consider two further non binary variants of this latter strategy. One referred to as *uniqueness constraint cost* **(UCC)**, that assumes as confidence the negative of c_1, and one referred to as *uniqueness constraint occurrences* **(UCO)**, that assumes that confidence is inversely proportional to the number of collisions. For the latter four outlined strategies the other candidates in the pool of colliding points are always set to invalid.

3.4. Disparity map features

Confidence measures belonging to this group are obtained by extracting features from the reference disparity map. Therefore they are potentially suited to infer confidence for any 3D sensing device. *Distance to discontinuity* **(DTD)** [41, 28] determines for each point the distance to the supposed closest depth boundary while, for the same purpose, *disparity map variance* **(DMV)** computes the disparity gradient module [8]. Remaining confidence measures belonging to this category extract features on a patch centered on the examined point. *Variance of disparity* **(VAR)** [28, 31] computes the disparity variance, *disparity agreement* **(DA)** [31] counts the number of points having the same disparity of the central one, *median deviation of disparity* **(MDD)** [41, 28, 31] computes the difference between disparity and its median and *disparity scattering* **(DS)** [31] encodes the number of different disparity assignments on the patch.

3.5. Reference image features

Confidence measures belonging to this category use as domain only the reference image. *Distance to border* **(DB)** [41, 28] aims at detecting invalid disparity assignments often originated in the image border due to the stereo setup. Assuming the left image as reference a more meaningful variant of DB, referred to as *distance to left border* **(DLB)**, deploys the distance to the left border. Both measures rely on prior information and not on image content. The last two confidence measure of this category extract features from the reference image: *horizontal gradient measure* **(HGM)** [8, 28] analyses the response to horizontal gradients in order to detect image texture while *distance to edge* **(DTE)** attempts to detect depth boundaries, sometimes unreliable for stereo algorithms, according to the distance to the closest edge.

3.6. Image distinctiveness

The idea behind these confidence measures is to exploit the notion of distinctiveness of the examined point within its neighborhoods along the horizontal scanline of the same image. *Distinctiveness* **(DTS)** [19, 13] exactly leverages on such definition by assuming as confidence for a given point the lowest *self-matching* cost computed within a certain prefixed range excluding the point under examination. *Distinctive similarity measure* **(DSM)** [45, 13] assigns as confidence value to a given point the product of two DTSs, one computed on the reference image and the other one on the right image in the location of the assumed homologous point, divided by the square of c_1 [13] or c_1 [45]. For a given point the *self-aware matching measure* **(SAMM)** [26, 13] computes the zero mean normalized correlation between the left-right cost curve, appropriately translated according to the assumed disparity, and the left-left cost curve.

3.7. Learning-based approaches

Recently, some authors proposed to infer confidence measures exploiting machine learning frameworks. A common trend in such approaches consists in feeding a random forest classifier with multiple confidence measures [8, 41, 28, 31] or deploying for the same purpose deep learning architectures [40, 32]. A notable difference with conventional confidence measures reviewed so far, is that learning-based approaches require a training phase, on datasets with ground-truth or by means of appropriate methodologies [27, 43], to infer the degree of uncertainty of disparity assignments.

3.7.1 Random forest approaches

In this category a seminal approach is represented by *ensemble learning* **(ENS_c)** [8]. This method infers a confidence measure by feeding to a random forest, trained for classification, a feature vector made of 23 confidence measures extracted from the original stereo pair, the left and right disparity maps and the cost volumes computed on the stereo pair at different scales. Then, the resulting features are up-sampled to the original resolution. The feature vector consists of the following measures: $\text{PKR}^{1,2,3}$, $\text{NEM}^{1,2,3}$, $\text{PER}^{1,2,3}$, LRC^{1}, $\text{HGM}^{1,2,3}$, $\text{DMV}^{1,2,3}$, $\text{DAM}^{1,2,3}$, $\text{ZSAD}^{1,2,3}$ and SGE^{1}. The superscript refers to the scale: 1 original resolution, 2 half-resolution and 3 quarter-resolution. The authors advocate to train the random-forest with such feature vector for classification *"as confidence measures do not contain matching error magnitude information"*, by extracting the posterior probability of the predicted class at inference time. However, the average response over all the trees in the forest can be used as well by training in *regression*. Therefore, we also include in our evaluation *ensemble learning in regression mode* **(ENS_r)** that to the best of our knowledge has not been considered before. In *ground control point* **(GCP) [41]** the confidence measure is inferred by feeding to a random forest, trained in regression mode, a feature vector containing 8 measures computed at the original scale. The features extracted from left image, left and right disparity maps and the cost volume are: MSM, DB, MMN, AML, LRC, LRD, DTD and MDD. In *leveraging stereo* **(LEV) [28]** a feature vector containing 22 measures extracted from the left image, left and right disparity maps and cost volume is fed to a random forest trained for regression. The feature vector, superscript encodes the patch size, consists of: PKR, PKRN, MSM, MM, WMN, MLM, PER, NEM, LRC, LRD, LC, DTD, $\text{VAR}^{1,2,3,4}$, $\text{MDD}^{1,2,3,4}$, HGM and DLB. Differently from previous approaches, *O(1) disparity features* **(O1) [31]** proposes a method entirely based on features extracted in constant time from the left disparity map. The feature vector, superscript encodes the patch size, consists of: $\text{DA}^{1,2,3,4}$, $\text{DS}^{1,2,3,4}$, $\text{MED}^{1,2,3,4}$, $\text{MDD}^{1,2,3,4}$ and $\text{VAR}^{1,2,3,4}$, being MED the median of disparity. As for ENS_r, GCP and LEV the feature vector is fed to a random forest trained in regression mode. We conclude this section observing that ENS [8] and LEV [28] also propose variants of the original method with a reduced number of features, respectively 7 and 8. For LEV, the features are selected analyzing the importance of variable once trained the random forest with the full 22 feature vector and then retraining the network. However, as reported in [8] and [28], being higher the effectiveness of full feature vectors, we consider in our evaluation such versions of ENS, in classification and regression mode, and LEV.

3.7.2 CNN approaches

As for many other computer vision fields, convolutional neural networks have recently proven to be very effective

also for confidence estimation. In *patch based confidence prediction* **(PBCP)** [40] the input of a CNN consists of two channels p_1 and p_2 computed, on a patch basis, from left and right disparity maps. Being patch values strictly related to their central pixel, confidence map computation is pretty demanding. A faster solution, made of patches no longer related to central pixels, allows for a very efficient confidence map prediction according to common optimization techniques in deep learning, with a minor reduction of effectiveness. However, being the full-version more effective we consider this one in our experiments.

A step towards a further abstraction is represented by *confidence CNN* **(CCNN) [32]**. In fact, in this approach confidence prediction is regressed by a CNN without extracting any cue from the input data. The deep network, trained on patches, learns from scratch a confidence measure by processing only the left disparity map. This property, shared with O1, makes these methods potentially suited to any 3D sensor [31, 32].

3.8. SGM specific

This category groups two approaches intrinsically related to SGM [9]. The idea behind these approaches is to exploit intermediate results available in such stereo algorithm to infer a confidence map. Specifically, the *local-global relation* **(PS)** [20] combines the cues available in the cost curve before and after semi-global optimization, while *sum of consistent scanlines* **(SCS)** [11] counts for each pixel the number of scanlines voting for the same disparity assigned by the full SGM pipeline.

4. Evaluation protocol and experimental results

In this section, we report exhaustive experimental results concerning different aspects related to the examined confidence measures on the following datasets K12 (194 images), K15 (200 images) and M14 (15 images). For each dataset we consider the stereo pairs belonging to the *training set* being the ground-truth available. We include in the evaluation all the measures previously reviewed including any variant. Moreover, for patch-based ones (*i.e.*, APKR, APKRN, WPKR, WPKRN, DA, DS, MED, VAR) we consider patches of different size (*i.e.*, 5×5, 7×7, 9×9 and 11×11 corresponding to superscript 1,2,3,4 in LEV and O1 features) being the scale effective according to [28, 31]. Of course, we consider state-of-the-art methods based on random forests, including variant ENS_r, and the two approaches based on CNNs. Overall, we evaluate 76 confidence measures[1]. In Section 4.1 we assess with three stereo algorithms the performance of such measures when dealing with the selection of correct matches by means of the ROC curve analysis proposed in [13] and widely adopted in this field [8, 41, 28, 31, 32, 40]. Moreover, since machine learning is the key technology behind most recent approaches, in Section 4.2 we report how training affects their effectiveness focusing in particular on the amount of training samples and the capability to generalize across different data (*i.e.*, datasets). Finally, being confidence measures often employed to improve stereo accuracy [41, 28, 31, 40], in Section 4.3 we assess the performance of the most effective confidence measures when plugged in one of such state-of-the-art methods [28].

[1]Source code available at vision.disi.unibo.it/~mpoggi/code.html

4.1. Detection of correct matches

The ability to distinguish correct disparity assignments from wrong ones is the most desirable property of a confidence measure. To quantitatively evaluate this, [13] adopted ROC curve analysis, measuring the capability of removing errors from a disparity map according to the confidence values. That is, given a disparity map, a subset p of pixels is extracted in order of decreasing confidence (*e.g.*, 5% of the total pixels) and the error rate on such sample is computed, as the percentage of points with an absolute distance from ground-truth value higher than a threshold τ, varying with the dataset. Then, the subset is increased by extracting more pixels (*e.g.*, an additional 5%) and the error rate is computed, until all the pixels in the image are considered. Ties are solved by including all the tying pixels in the subsample. The relation between each sub-sample p and its error rate draws a ROC curve and its AUC measures the capability of the confidence measure to effectively distinguish good matches from wrong ones. Considering a disparity map with a portion $\varepsilon \in [0, 1]$ of erroneous pixels, an optimal measure would be able to achieve a 0 error rate when extracting the first $(1 - \varepsilon)$ points. Thus, the optimal AUC value [13] can be obtained as follows

$$AUC_{opt} = \int_{1-\varepsilon}^{\varepsilon} \frac{p - (1-\varepsilon)}{p} dp = \varepsilon + (1-\varepsilon)\ln(1-\varepsilon) \tag{1}$$

Following this protocol, we evaluate the 76 confidence measures on K12, K15 and M14 with three popular stereo algorithms adopting the *winner takes all* strategy for disparity selection:

- AD-CENSUS: aggregates matching costs, computed on 5×5 patches with census transform [46], with a 5×5 box-filter.

- MC-CNN [47]: local method inferring costs from image patches using a CNN. We used the same networks trained by the authors on K12, K15 and M14.

(a)

	K12 (ε = 38.82%)			K15 (ε = 35.41%)			M14 (ε = 37.78%)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	4^{12}	0.1806	$APKR_{11}$	4^{12}	0.1541	$APKR_{11}$	4^{7}	0.1355
3.2	WMNN	7^{34}	0.2215	WMN	7^{34}	0.2024	WMN	6^{23}	0.1579
3.3	LRD	5^{20}	0.1946	LRD	6^{28}	0.1825	LRD	5^{21}	0.1519
3.4	DA_{11}	3^{8}	0.1668	DA_{11}	3^{7}	0.1399	DA_{11}	3^{4}	0.1294
3.5	DB	8^{65}	0.3446	DB	8^{66}	0.3103	DLB	8^{69}	0.3333
3.6	SAMM	6^{25}	0.2030	SAMM	5^{20}	0.1715	DSM	7^{40}	0.1798
3.7.1	O1	2^{3}	0.1309	O1	2^{3}	0.1128	O1	2^{3}	0.1211
3.7.2	**CCNN**	$\mathbf{1}^{1}$	**0.1223**	**CCNN**	$\mathbf{1}^{1}$	**0.1041**	**CCNN**	$\mathbf{1}^{1}$	**0.1128**
Optimal			0.1067			0.0884			0.0899

(b)

Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	7	11	44
ENS_r	5	5	33
GCP	6	6	8
LEV	4	4	5
O1	3	3	3
PBCP	2	2	2
CCNN	**1**	**1**	**1**

(c)

	K12 (ε = 17.10%)			K15 (ε = 15.37%)			M14 (ε = 26.70%)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	4^{11}	0.0566	$APKR_{11}$	4^{11}	0.0508	$APKR_{11}$	3^{5}	0.0728
3.2	WMN	6^{30}	0.0748	WMN	6^{31}	0.0654	WMN	4^{13}	0.0763
3.3	LRD	7^{31}	0.0748	LRD	7^{32}	0.0712	UCC	5^{22}	0.0896
3.4	DS_9	3^{8}	0.0542	DS_9	3^{8}	0.0477	DS_{11}	6^{35}	0.1061
3.5	DLB	8^{66}	0.1543	HGM	8^{67}	0.1439	DLB	8^{68}	0.2260
3.6	SAMM	5^{16}	0.0598	SAMM	5^{21}	0.0557	DSM	7^{40}	0.1228
3.7.1	O1	2^{2}	0.0317	O1	2^{2}	0.0324	O1	2^{3}	0.0680
3.7.2	**CCNN**	$\mathbf{1}^{1}$	**0.0297**	**CCNN**	$\mathbf{1}^{1}$	**0.0297**	**CCNN**	$\mathbf{1}^{1}$	**0.0637**
Optimal			0.0231			0.0213			0.0459

(d)

Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	7	7	24
ENS_r	5	5	17
GCP	6	6	14
LEV	4	4	4
O1	2	2	3
PBCP	3	3	2
CCNN	**1**	**1**	**1**

(e)

	K12 (ε = 16.78%)			K15 (ε = 13.68%)			M14 (ε = 25.91%)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	3^{7}	0.0492	$APKR_{11}$	3^{7}	0.0457	$APKR_{9}$	2^{2}	0.0739
3.2	WMN	4^{11}	0.0554	WMN	5^{12}	0.0502	WMN	4^{8}	0.0.779
3.3	UCC	6^{21}	0.0735	UCC	6^{19}	0.0640	UCC	6^{23}	0.0959
3.4	DS_{11}	5^{12}	0.0554	DS_{11}	4^{11}	0.0501	DS_{11}	5^{13}	0.0884
3.5	DB	9^{67}	0.1378	DB	9^{68}	0.1265	DLB	9^{70}	0.2157
3.6	DSM	7^{36}	0.0811	DSM	7^{28}	0.0679	DSM	7^{32}	0.1041
3.7.1	LEV	2^{2}	0.0358	O1	2^{2}	0.0323	O1	3^{6}	0.0777
3.7.2	**CCNN**	$\mathbf{1}^{1}$	**0.0358**	**CCNN**	$\mathbf{1}^{1}$	**0.0302**	**CCNN**	$\mathbf{1}^{1}$	**0.0736**
3.8	SCS	8^{41}	0.0851	SCS	8^{48}	0.0790	SCS	8^{36}	0.1080
Optimal			0.0227			0.0184			0.0431

(f)

Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	27	31	44
ENS_r	5	5	11
GCP	6	6	28
LEV	2	4	19
O1	3	2	6
PBCP	4	3	7
CCNN	**1**	**1**	**1**

Table 1. Detection of correct matches with three stereo algorithms - top (a,b) AD-CENSUS, middle (c,d) MC-CNN and bottom (e,f) SGM - and three datasets K12, K15 and M14. For each algorithm there are two tables. On the left the best confidence measure for each category (*e.g.*, 3.1 refers to measures belonging to the category reviewed in Section 3.1), the ranking (within categories and, in superscript, absolute) and the AUC. On the right, the absolute ranking of learning-based confidence measures. We also report average error rate ε for each dataset on the top labels. Concerning categories 3.7.1 and 3.7.2 we trained each confidence measure on the first 20 images of K12 with the considered algorithm (*i.e.*, (a,b) with AD-CENSUS, (c,d) with MC-CNN and (e,f) with SGM).

- SGM [9]: eight scanline implementation with AD-CENSUS aggregated costs as data term and P1 and P2, respectively, 0.2 and 0.5 (being costs normalized).

Concerning confidence measures based on machine learning, for each stereo algorithm, we train each one on a subset of images from the K12 dataset (the first 20 images, extracting a sample from each pixel with available ground-truth, for a total of 2.7 million samples) and evaluate it on all the datasets (for K12 excluding the training images), in order to assess their performance on very different scenes. For approaches based on random forests we train on 10 trees as suggested in [28] and adopting a fixed number of iteration as termination criteria (*e.g.*, proportional to the number of trees), while we train CNN based measures for 25 epochs (resulting in about 1 million iterations), with a batch of size 64, *learning rate* of 0.001 and *momentum* of 0.9, by minimizing the loss functions reported in [32, 40]. Different training sets (*e.g.*, datasets, number of samples and so on) may lead to different performance. This fact will be thoroughly evaluated in Section 4.2. For the evaluation reported in this section we trained only on K12 in order to assess how much a confidence measure is able to generalize its behavior across different datasets which is an important and desirable feature in most practical applications. We adopt as error bound $\tau = 3$ for K12 and K15 and $\tau = 1$ for M14[2] as suggested in the corresponding papers.

In Table 1 we summarize results in terms of AUC averaged on each dataset (K12, K15 and M14) for AD-CENSUS (a,b), MC-CNN (c,d) and SGM (d,e), reporting the aver-

[2]Middlebury frames have been processed at quarter resolution to level out the original disparity range with other datasets (800 vs 228 for KITTIs).

age error rate ε for each dataset. For each algorithm we report on the left table the best measure for each category described in Section 3 and its absolute ranking and, on the right table, the absolute ranking for confidence measures based on machine learning. Observing tables 1 (a,c,e), we can notice that these latter measures always yield the best results, with CCNN systematically the top-performing one in terms of AUC, and the ones based on random forest following very close (with O1 the best in its category in 7 out of 9 experiments). Focusing on categories 3.7.1 and 3.7.2, we can notice that in most cases PBCP, O1 and LEV perform very well with the exception of the SGM algorithm and M14 (Table 1(f)). In this specific case, excluding CCNN, APKR$_{11}$ performs better than approaches based on machine learning. Anyway, in this case too, the effectiveness of O1 and PBCP seems acceptable. This fact highlights that some confidence measure based on learning approaches (in particular CCNN but also O1 and PBCP) have excellent performance across different data. Interestingly, such measures use as input cue only the disparity maps. Tables 1 (b,d,f) also show that for other measures such as ENS$_c$, ENS$_r$, GCP and LEV this behavior is not always verified, in particular with M14. Finally, we observe that ENS$_r$ always (and sometimes significantly) outperforms ENS$_c$. Concerning other categories, we can notice that APKR yields good results in all the experiments and not only with M14 and SGM as already highlighted. Other interesting confidence measures are those belonging to category 3.4 and in particular DA with AD-CENSUS and DS with MC-CNN and SGM. Such results confirm that processing cues from the disparity map only, as done by best learning-based approaches, yields reliable confidence estimation. Other categories do not seem particularly effective, especially those based only on left image cues have always the overall worst performance. For measures belonging to category 3.2, though not very effective excluding experiments with SGM, WMN always achieves the best results. Besides, it's worth pointing out that naive versions of traditional strategies produce worse AUC values than their original counterparts. Regarding SGM-specific methods, SCS always outperforms PS but with AUC values quite far from the top-performing approaches. Finally, concerning categories 3.3 and 3.6, such measures on the three datasets do not grant reliable confidence prediction.

4.2. Impact of training data

Having assessed the performance of the confidence measure with different algorithms and datasets, this section aims at analyzing the impact of training data on the effectiveness of learning-based measures. To quantitatively compare the results between different training configuration, we define Δ_k as the ratio between the AUC value achieved by the measure k and the AUC$_{opt}$ as,

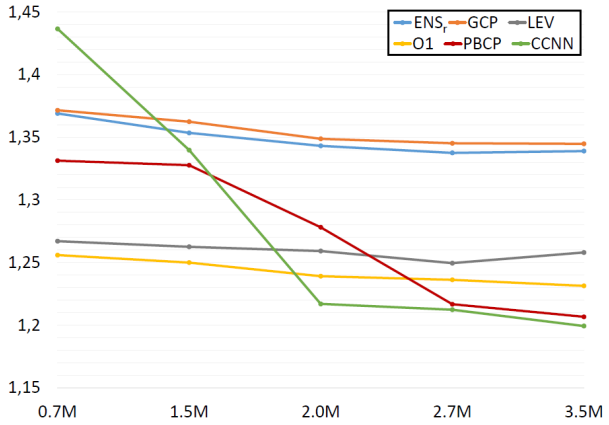

Figure 1. Ratio between the average AUC achieved by learning-based confidence measures trained with different number of samples from K12 and the optimal AUC. Evaluated on the rest of K12 with AD-CENSUS algorithm.

$$\Delta_k = \frac{AUC_k}{AUC_{opt}} \tag{2}$$

The lower the Δ_k, the better the training configuration.

The first issue we are going to evaluate is the amount of training samples required and how it affects the overall effectiveness of each confidence measure. We carried out multiple trainings with a different number of samples obtained from 5, 10, 15, 20 and 25 stereo pairs of K12 dataset starting from the first image. These subsets provide, respectively, about 0.7, 1.5, 2, 2.7 and 3.5 million samples with available ground-truth for training. By using more data we can deploy more complex random forests as well. Nevertheless, we keep the same parameters and termination criteria described in Section 4.1 to compare the behavior of the same forest fed with different feature vectors when more samples are available. Figure 1 reports Δ_k, as a function of the number of training samples, for the best six measures based on machine learning (*i.e.*, ENS$_r$, GCP, LEV, O1, CCNN and PBCP) trained on AD-CENSUS algorithm. We can notice how the amount of training data slightly changes the effectiveness of the methods based on random forest (less than 0.05 Δ_k improvement), highlighting how the best AUC is obtained starting from 2.7 million samples. Conversely, measures based on CNNs improve their effectiveness by a significant margin only when trained on a sufficiently larger amount of data, but such improvement almost saturates at 2.7 million samples. In particular, we can observe how CCNN achieves the worst results when trained with the smallest subset of images, resulting to be the best measure with a larger training set (with a Δ_k margin of about 0.25). Excluding LEV and ENS$_r$ at 3.5M, all the measures show a monotonic improvement in terms of AUC by increasing the number of samples.

The second issue evaluated concerns how much a con-

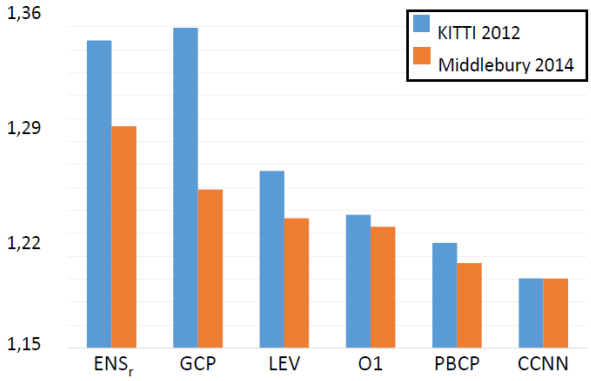

Figure 2. Experimental results on M14. Ratio between the average AUC achieved by each confidence measure, trained on K12 (blue) and M14 (orange), and the optimal AUC evaluated on the rest of M14 with AD-CENSUS algorithm.

	K12		K15		M14	
	bad3	avg	bad3	avg	bad1	avg
SGM	16.53	7.40	13.68	6.13	25.91	7.11
APKR$_{11}$	11.26^{10}	3.60^{10}	9.57^{10}	2.94^{10}	23.79^{8}	5.15^{10}
SAMM	10.95^{6}	3.15^{6}	9.13^{6}	2.58^{6}	24.07^{10}	4.94^{4}
DA$_{11}$	11.18^{9}	3.40^{9}	9.50^{9}	2.77^{9}	23.98^{9}	5.10^{9}
ENS$_c$	10.42^{2}	2.71^{4}	9.02^{4}	2.33^{4}	23.49^{4}	5.00^{8}
ENS$_r$	10.63^{5}	2.95^{5}	9.08^{5}	2.46^{5}	23.74^{7}	4.96^{6}
GCP	11.05^{8}	3.26^{8}	9.28^{7}	2.67^{7}	23.54^{5}	4.97^{7}
LEV	10.97^{7}	3.22^{7}	9.34^{8}	2.72^{8}	23.67^{6}	4.94^{5}
O1	**10.41**^{1}	**2.36**^{1}	8.79^{2}	1.84^{2}	23.18^{3}	4.07^{2}
PBCP	10.63^{4}	2.60^{3}	8.86^{3}	1.91^{3}	22.92^{2}	**3.95**^{1}
CCNN	10.61^{3}	2.41^{2}	**8.79**^{1}	**1.80**^{1}	**22.86**^{1}	4.12^{3}

Table 2. Error rate (percentage) and average pixel error on the three datasets achieved by vanilla SGM (first row) and the confidence modulation proposed in [28] plugging: APKR$_{11}$, SAMM, DA$_{11}$, ENS$_c$, ENS$_r$, GCP, LEV (the one proposed in [28]), O1, PBCP and CCNN. Learning-based confidence measures trained, with AD-CENSUS, on the first 20 images of K12.

fidence measure can generalize across different environments/scenes (*i.e.*, datasets). To quantitatively evaluate this behavior, we trained with AD-CENSUS the confidence measures on a subset of M14, processing an almost equivalent amount of training samples with respect to the training configuration adopted in Section 4.1. Then, we compared the results achieved with this configuration to the one used in Section 4.1 with AD-CENSUS on the remaining data from M14, computing Δ_k as defined in Equation 2. A confidence measure achieving similar Δ_k in the two configuration is able to generalize well between the two very different scenarios. Figure 2 plots the two values for the six confidence measures. We can clearly notice how measures based on CNNs better generalize with respect to random forest approaches, with CCNN being more effective in this sense than PBCP. Moreover, O1 appears to better adapt to different data, achieving a lower margin between the two Δ_k with respect to ENS$_r$, GCP and LEV. This experiment highlights once again that confidence measures using as input cue the disparity map(s) (*i.e.*, CCNN, PBCP and O1) seem less prone to under-fitting.

4.3. Improvements to stereo accuracy

The final issue we investigated is the impact of confidence measures on stereo accuracy, a topic that recently gained a lot of attention (*e.g.*, [41, 28, 31, 40]). For this evaluation we choose the cost modulation proposed by Park and Yoon [28]. The reason is that differently from [31], which is specific for SGM algorithm, and [41, 40], based on parameters potentially different from measure to measure, [28] is suited for any stereo algorithm and parameter-free. SGM was tuned as reported in Section 4.1. We plugged in [28] the machine learning based measures, as well as three standalone measures (*i.e.*, APKR, SAMM and DA$_{11}$). On the three datasets K12, K15 and M14, from Table 2 we can notice that confidence measures based on machine learning are overall more effective than other ones. In particular, O1 achieves the lowest error rate with K12 and CCNN and PBCP outperforms other ones in K15 and M14. This experiment highlights that there is not a direct relationship with the effectiveness of the confidence measure in terms of AUC. However, most effective confidence measures (*i.e.*,, CCNN, PBCP and O1) according to this metric achieve the best results. Finally we point out that in this experiments, ENS$_c$ and ENS$_r$, frequently perform better than others confidence measures, conventional and learning-based ones. Moreover, for their deployment in cost modulation ENS$_c$ outperforms ENS$_r$ most of the times, conversely to what observed in terms of AUC.

5. Conclusions

In this paper we have reviewed and evaluated state-of-the-art confidence measures focusing our attention on recent ones based on machine learning techniques. Our exhaustive evaluation, with three stereo algorithms and three large and challenging datasets, clearly highlights that learning-based ones are much more effective than conventional approaches. In particular, those using as input cue the disparity maps achieve better results in terms of detection of correct match, capability to adapt to new data and effectiveness to improve stereo accuracy. In such methods training is certainly an additional issue but, as reported in our evaluation, the overall amount of training data required is limited and best learning-based confidence measures much better generalize to new data.

Acknowledgement

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

References

[1] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. 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

[2] L. Di Stefano, M. Marchionni, and S. Mattoccia. A fast area-based stereo matching algorithm. *Image and vision computing*, 22(12):983–1005, 2004. 2, 3

[3] G. Egnal, M. Mintz, and R. P. Wildes. A stereo confidence metric using single view imagery. In *PROC. VISION INTERFACE*, pages 162–170, 2002. 3

[4] F. Garcia, B. Mirbach, B. E. Ottersten, F. Grandidier, and I. Cuesta-Contreras. Pixel weighted average strategy for depth sensor data fusion. In *ICIP*, pages 2805–2808. IEEE, 2010. 2

[5] 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. 1, 2

[6] R. Gherardi. Confidence-based cost modulation for stereo matching. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, Dec 2008. 2

[7] R. Haeusler and R. Klette. Evaluation of stereo confidence measures on synthetic and recorded image data. In *2012 International Conference on Informatics, Electronics and Vision, ICIEV 2012*, pages 963–968, 2012. 2, 3

[8] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR. Proceedings*, pages 305–312, 2013. 1. 1, 2, 3, 4, 5

[9] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):328–341, feb 2008. 1, 2, 3, 5, 6

[10] H. Hirschmüller, P. R. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *Int. J. Comput. Vision*, 47(1-3), apr 2002. 3

[11] H. Hirschmuller, M. Buder, and I. Ernst. Memory efficient semi-global matching. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 371–376, 2012. 5

[12] H. Hirschmller. Evaluation of cost functions for stereo matching. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. 2

[13] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 2121–2133, 2012. 1, 2, 3, 4, 5

[14] 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. 2, 3

[15] S. Kim, C. Y. Jang, and Y. H. Kim. Weighted peak ratio for estimating stereo confidence level using color similarity. In *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 196–197, Oct 2016. 2, 3

[16] D. Kong and H. Tao. A method for learning matching errors in stereo computation. In *British Machine Vision Conference (BMVC)*, 2004 2004. 2

[17] S. Lefebvre, S. Ambellouis, and F. Cabestaing. A colour correlation-based stereo matching using 1D windows. In IEEE, editor, *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, SITIS'07*, pages 702–710, Shanghai, China, Dec 2007. IEEE. 3

[18] W. Luo, A. G. Schwing, and R. Urtasun. Efficient Deep Learning for Stereo Matching. In *Proc. CVPR*, 2016. 2

[19] R. Manduchi and C. Tomasi. Distinctiveness maps for image matching. In *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pages 26–31. IEEE, 1999. 4

[20] G. Marin, P. Zanuttigh, and S. Mattoccia. Reliable fusion of tof and stereo depth driven by confidence measures. In *14th European Conference on Computer Vision (ECCV 2016)*, pages 386–401, 2016. 2, 5

[21] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *Int. J. Comput. Vision*, 8(1), jul 1992. 3

[22] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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. 1, 2

[23] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[24] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. M. Frahm, R. Yang, D. Nister, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007. 2, 3

[25] D. B. Min and K. Sohn. An asymmetric post-processing for correspondence problem. *Sig. Proc.: Image Comm.*, 25(2):130–142, 2010. 2, 3

[26] P. Mordohai. The self-aware matching measure for stereo. In *The International Conference on Computer Vision (ICCV)*, pages 1841–1848. IEEE, 2009. 2, 4

[27] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Using self-contradiction to learn confidence measures in stereo vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4

[28] M. G. Park and K. J. Yoon. Leveraging stereo matching with learning-based confidence measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2, 3, 4, 5, 6, 8

[29] D. Pfeiffer, S. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *IEEE Computer Vision and Pattern Recognition*, pages 297–304, Portland, OR, USA, June 2013. 2

[30] 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. 2

[31] 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. 1, 2, 3, 4, 5, 8

[32] M. Poggi and S. Mattoccia. Learning from scratch a confidence measure. In *Proceedings of the 27th British Conference on Machine Vision, BMVC*, 2016. 1, 2, 4, 5, 6

[33] M. Poggi and S. 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. 2

[34] M. Poggi, F. Tosi, and S. Mattoccia. Efficient confidence measures for embedded stereo. In *19th International Conference on Image Analysis and Processing (ICIAP 2017)*, September 2017. 2

[35] M. Poggi, F. Tosi, and S. Mattoccia. Even more confident predictions with deep machine-learning. In *12th IEEE Embedded Vision Workshop (EVW2017) held in conjunction with IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[36] N. Sabater, A. Almansa, and J. M. Morel. Meaningful Matches in Stereovision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(5):930–42, dec 2011. 2

[37] D. Scharstein, H. Hirschmller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, pages 31–42. 1, 2

[38] D. Scharstein and R. Szeliski. Stereo matching with non-linear diffusion. *International Journal of Computer Vision*, 28:155–174, 1998. 3

[39] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 47(1-3):7–42, apr 2002. 1, 2

[40] A. Seki and M. Pollefeys. Patch based confidence prediction for dense disparity map. In *British Machine Vision Conference (BMVC)*, 2016. 1, 2, 4, 5, 6, 8

[41] A. Spyropoulos, N. Komodakis, and P. 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. 1, 2, 3, 4, 5, 8

[42] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 24-26 June 2008, Anchorage, Alaska, USA*, 2008. 2

[43] F. Tosi, M. Poggi, A. Tonioni, L. Di Stefano, and S. Mattoccia. Learning confidence measures in the wild. In *28th British Machine Vision Conference (BMVC 2017)*, September 2017. 2, 4

[44] A. Wedel, A. Meiner, C. Rabe, U. Franke, and D. Cremers. Detection and Segmentation of Independently Moving Objects from Dense Scene Flow. In *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 14–27, Bonn, Germany, August 2009. Springer. 3

[45] K.-J. Yoon and I.-S. Kweon. Distinctive similarity measure for stereo matching under point ambiguity. *Computer Vision and Image Understanding*, 112(2):173–183, 2008. 4

[46] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the Third European Conference on Computer Vision (Vol. II)*, ECCV '94, pages 151–158, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc. 5

[47] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 1, 2, 5