Beyond local reasoning for stereo confidence estimation with deep learning

Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia

University of Bologna, Viale del Risorgimento 2, Bologna, Italy
{fabio.tosi5,m.poggi,stefano.mattoccia}@unibo.it
http://vision.disi.unibo.it/~ftosi{mpoggi,smatt}

Abstract. Confidence measures for stereo gained popularity in recent years due to their improved capability to detect outliers and the increasing number of applications exploiting these cues. In this field, convolutional neural networks achieved top-performance compared to other known techniques in the literature by processing local information to tell disparity assignments from outliers. Despite this outstanding achievements, all approaches rely on clues extracted with small receptive fields thus ignoring most of the overall image content. Therefore, in this paper, we propose to exploit nearby and farther clues available from image and disparity domains to obtain a more accurate confidence estimation. While local information is very effective for detecting high frequency patterns, it lacks insights from farther regions in the scene. On the other hand, enlarging the receptive field allows to include clues from farther regions but produces smoother uncertainty estimation, not particularly accurate when dealing with high frequency patterns. For these reasons, we propose in this paper a multi-stage cascaded network to combine the best of the two worlds. Extensive experiments on three datasets using three popular stereo algorithms prove that the proposed framework outperforms state-of-the-art confidence estimation techniques.

Keywords: confidence measures, stereo matching, deep learning

1 Introduction

Stereo is a popular technique to infer the 3D structure of a scene sensed by two cameras and for this reason deployed in several computer vision applications. A stereo setup is typically made of two synchronized cameras and establishing correspondences between homologous points allows inferring depth through simple triangulation. Consequently, stereo literature is extremely vast since it dates back to the '60s and since then has been very popular. Despite this longstanding research activity, due to its ill-posed nature, algorithms aimed at finding stereo correspondences may lead to inaccurate results. In particular, when dealing with occlusions, transparent or reflecting surfaces, texture-less regions. Thus, on the one hand, we need accurate depth estimation algorithms. On the other hand,

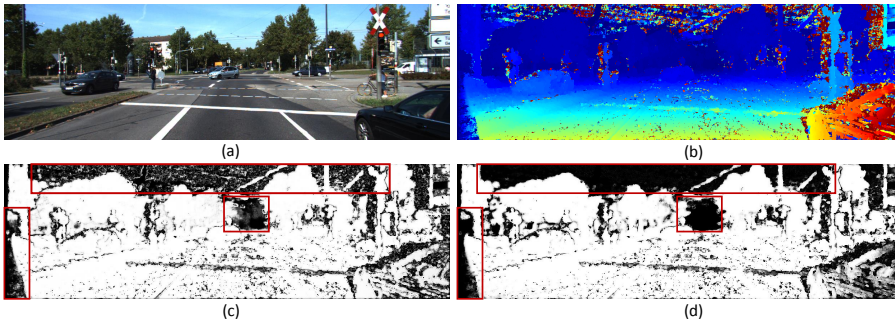

Fig. 1. Example of confidence estimation. (a) Reference image from KITTI 2015 dataset [7], (b) disparity map obtained with MC-CNN [8], (c) confidence estimated with a local approach (CCNN [2]) and (d) the proposed local-global framework, highlighting regions on which the latter method provides more reliable predictions (red bounding boxes).

given a depth or disparity map, we need an accurate methodology to infer the degree of reliability of each point. This task is referred to as confidence estimation and is of paramount importance when dealing with depth data.

Among the many confidence estimators proposed in the literature, recently reviewed and evaluated by Poggi et al. in [1], methods using as input cue information extracted from the disparity domain only [2,3,4] proved to be particularly effective. Compared to approaches relying on cues extracted from the cost volume or other strategies known in the literature, these methods currently represent state-of-the-art. Another notable advantage of methods working in the disparity domain, and in particular [2,4], is their ability to cope with depth data inferred by stereo systems not exposing to the user the cost volume, such as those based on closed source software or commercial stereo cameras. Regardless of this fact, machine learning deeply impacted confidence estimation starting from the seminal work of Haeusler et al. [5] aimed at inferring a confidence measure combining conventional confidence measure within a random forest framework. Later, other works successfully followed this strategy and, more recently, methods based on Convolutional Neural Networks (CNNs) achieved outstanding results [1] by inferring a confidence score for each pixel of a disparity map feeding to the deep-network a patch centered on it. In contrast to CNN-based method [3] and approaches based on random-forest, CCNN [2] accomplishes this task without relying on any hand-crafted feature defined beforehand. Currently, CCNN represents state-of-the-art for confidence estimation as recently highlighted in [1]. This strategy was extended in [6] by feeding to the CNN also the input reference image with promising results deploying, however, a larger amount of training samples.

Regardless of the strategy adopted, all these methods estimate confidence with a relatively small receptive field intrinsic in their local patch-based nature. Increasing such parameter in these methods does not enable significant improve-

ments and may also lead to poor results. Thus, state-of-the-art methods do not take advantage of the whole image and disparity content. Although this strategy is undoubtedly valid, on the other hand, it seems clear that by looking at the whole reference image and disparity map matters for uncertainty estimation. This fact can be readily perceived by observing Figure 1 in the highlighted areas.

In particular, considering more global reasoning on the whole image and disparity content can improve the prediction for disparity values more unlikely to occur (e.g., objects extremely close to the camera), at the cost of a smoother prediction. This task can be undertaken by architectures with a large receptive field such as encoder-decoder models thus less accurate in the presence of high-frequency noise (e.g., outliers on the output of stereo algorithms such as AD-CENSUS or other matching functions). On the other hand, networks working on patches detect very well this kind of outliers but they are not able to capture farther information.

Therefore, in this paper, we propose to overcome this limitation by combining the best of the two worlds (i.e., networks based on small and large receptive fields). We do this by deploying a CNN-based architecture able to extract nearby and far-sighted cues, in the RGB and disparity domains, and to merge them to obtain a more accurate confidence estimation. By training a multi-modal cascaded architecture we first obtain two confidence predictions by reasoning respectively on local and farther cues, then we further elaborate on them to obtain a final, more accurate prediction. Figure 1 shows qualitatively how this strategy enables to estimate more reliable confidence scores.

To the best of our knowledge, our proposal is the first one enabling to i) exploit more global context for learning confidence predictions and ii) combine this novel technique with local approaches to design an effective local-global confidence measure. From now on, we will define as *global*, with abuse of language, a strategy going beyond traditional neighboring boundaries usually adopted in the field of confidence estimation. We extensively evaluate the proposed framework on three popular datasets, KITTI 2012 [9], KITTI 2015 [7] and Middlebury v3 [10] using three popular algorithms used in this field, respectively, AD-CENSUS [11], MC-CNN-fst matching cost [8] and SGM [12]. Such exhaustive evaluation clearly highlights that our proposal is state-of-the-art.

2 Related work

In this section, we review the literature concerning confidence measures, their applications and the most recent advances in stereo matching using deep learning being all these fields relevant to our proposal.

Confidence measures for stereo. Confidence measures have been extensively reviewed by Hu and Mordohai [13] and by Poggi et al. [1] more recently including methods based on machine-learning. While the first review evaluated confidence measures with standard local algorithm using *sum of absolute differences* (SAD) and *normalized cross correlation* (NCC) as matching costs on the Middlebury 2002 dataset [14], the second review considers recent state-of-

the-art confidence measures and evaluates them with three popular algorithms (AD-CENSUS [11], MC-CNN [8] and SGM [12]) on KITTI 2012 [9], KITTI 2015 [7] and Middlebury v3 [10] the standard datasets in this and other related fields. Both works follow the evaluation protocol defined in [13], consisting in Area Under the Curve (AUC) analysis from ROC curves. As reported in [1], machine learning enables to obtain more accurate confidence estimation compared to *conventional* strategies. Starting from the seminal work of Hausler et al. [5], other approaches fed hand-crafted features to a random forest classifier [5,15,16,4]. Recently, more accurate confidence estimators were obtained by leveraging on CNNs. In CCNN [2] Poggi and Mattoccia trained the network with raw disparity maps of the reference image while in PBCP [3] Seki and Pollefeys trained the network with pre-processed disparity maps concerned with reference and target images. According to the extensive evaluation reported in [1] both latter methods, and in particular CCNN, outperform any other known confidence measure. Poggi and Mattoccia [17] also proposed an effective strategy to improve confidence measures by exploiting local consistency. In [18] was proposed a method to improve random forest-based approaches for confidence fusion [15,16,4] by using a CNN. Fu et al. [6] extended CCNN [2] by adding the raw RGB image as input to the CCNN network. This strategy improves the final prediction when training on a much larger amount of training data (94 stereo pairs vs 20 images typically deployed with CCNN as in [1]). Some works looked deeper into the learning process of confidence measures, by studying features augmentation [19] or by designing self-supervised techniques to train them on static video sequences [20] or stereo pairs [21]. The latter technique proved to be effective even with CNN-based confidence measure CCNN. Finally, in [22] was proposed an evaluation of conventional confidence measures and their simplifications when targeting embedded systems.

Applications of confidence measures. While traditionally confidence measures were used to filter out outliers from disparity maps, some higher-level applications leveraging on them for other purposes have been deployed in the last years. Spyropoulos and Mordohai [15] used estimated confidence to detect very reliable disparity assignments (i.e., Ground Control Points) and setting for them ideal cost curves to improve the results of a further global optimization step. Park and Yoon [16] proposed a cost modulation function based on confidence applied to intermediate DSI (*Disparity Space Image*) before SGM optimization, Poggi and Mattoccia [4] modified the SGM pipeline to reduce the streaking effects along each scanline by penalizing low confidence hypothesis. Seki and Pollefeys [3] acted on P1 and P2 penalties of SGM tuning them according to the estimated confidence. In addition to these approaches, acting inside stereo algorithms to improve their final output, other applications concern sensor fusion [23] and disparity map fusion [24]. Shaked and Wolf [25] embedded confidence estimation inside a deep model stereo matching. Finally, confidence measures were also deployed for unsupervised adaptation of deep models for stereo matching [26] or unsupervised training of machine learning based measures [21], thus not requiring difficult to source disparity ground-truth labels.

Deep learning for stereo matching. The very first attempt to use deep learning in stereo matching was proposed in the seminal work of Zbontar and LeCun [27] aimed at inferring matching cost through a CNN by processing images patches. This technique, known as MC-CNN, is now deployed by many stereo pipelines as reported on the KITTI and Middlebury v3 benchmarks. By working on small image patches only (i.e., 9×9), deep learning based confidence measures [2,3,6] are affine to this approach, being all these methods based on small receptive fields. Recent advances in stereo consist of deploying deep networks embedding all steps of traditional pipelines. These models are typically characterized by encoder-decoder architectures, enabling an extremely large receptive field and thus able to incorporate most of the global image content. The first, seminal work in this direction is DispNet by Mayer et al. [28], followed more recently by GC-Net [29] and CLR [30].

Thus, although deep learning confidence measures working on image patches have been successfully proposed [2,3,6], the literature lacks global approaches for this task. Therefore, inspired by successful attempts based on encoder-decoder architectures for disparity estimation [28,30,29] and local approaches for confidence estimation, in this paper we combine both strategies to achieve a more robust confidence measure by exploiting cues inferred from local and global contexts.

3 Method overview

In this section, we introduce our local-global framework for confidence estimation. Driven by the recent success of confidence measures obtained by processing cues in the disparity domain only, and in particular those based on deep learning [2,3,6], we look beyond the small local neighborhood taken into account for each pixel by these methods and we analyze global context from both RGB and disparity domains to obtain a more consistent confidence estimation. Being local and global approaches characterized by complementary strengths, respectively the formers are very effective at detecting high-frequency patterns while the latter can incorporate much more cues from the surrounding pixels, we argued that combining them can further improve confidence estimation by overcoming the specific limitations of the single approaches. To do so, we will deploy two main architectures, respectively in charge of process local and global context. Then, the output of these two networks is combined to obtain the final prediction. In Section 3.1 we describe the local network, for which we choose state-of-the-art CCNN measure [2] and its extensions proposed in [6]. In Section 3.2 we introduce a novel architecture for *global* confidence estimation referred to as *ConfNet*, inspired by works concerning end-to-end stereo matching [28]. Finally, in Section 3.3 we outline our overall local-global framework combining cues generated by local and global approaches.

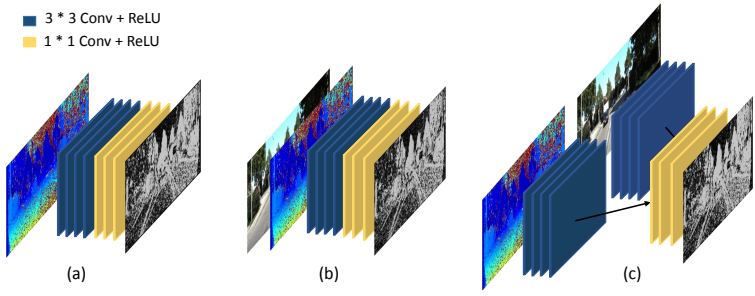

Fig. 2. Local architectures, respectively (a) CCNN [2], (B) EFN [6] and (c) LFN [6]. The networks uses 3×3 (blue) and 1×1 convolutional layers, all followed by ReLUs except the last one.

3.1 Local approaches

With local approaches, we refer to methodologies aimed at estimating the confidence score for a single pixel by looking at nearby pixels laying on a small local neighborhood. PBCP [3], CCNN [2] and multi-modal approaches [6] belongs to this category. We use the two latter techniques in our framework, depicted in Figure 2, because of the superior outliers detection performance achieved by the first [1] further improved, in some circumstances, by multi-modal networks [6]. Another reason to use CCNN-based networks is that both can be computed without requiring the right disparity map, required by PBCP [3], not always available in some circumstances as previously highlighted.

CCNN. This confidence measure is obtained by processing the disparity map through a shallow network, made of 4 convolutional layers with 3×3 kernels producing 64 features map at each level, followed by 2 convolutional layers with 1×1 kernels producing 100 features map and a final 1×1 convolution followed by Sigmoid activation to obtain confidence scores in $[0, 1]$ interval. All the other layers are followed by ReLU non-linearities. The first 4 layers do not apply any explicit padding to its input, thus reducing input size by 2 pixels on both height and width (i.e., 1 pixel on each side). This makes the single pixel confidence prediction bound to a 9×9 local patch, the receptive field of the network, centered on it. The fully convolutional nature of this model allows for training on image patches and then performs a single forward of a full resolution disparity map if properly padded (i.e., applying 4 pixel padding on each side).

Multi-modal networks. In [6] the authors propose to improve CCNN [2] by feeding to the network additional information from the RGB reference image. To this aim Fu et al. propose two strategies, respectively, the Early Fusion Network (EFN) and the Late Fusion Network (LFN). In the EFN, RGB and disparity patches are concatenated to form a 4-channel input, processed by a shallow network with the same structure of CCNN, but different number of channels at each

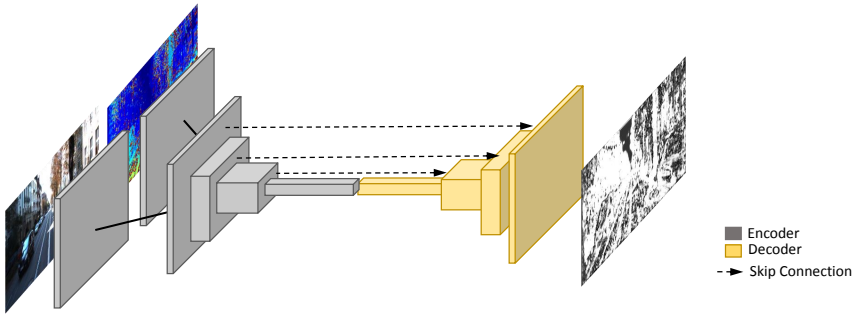

Fig. 3. ConfNet architecture. Encoding blocks (light blue) are made by 3×3 convolutions followed by batch normalization, ReLU and max-pooling. Decoding blocks (yellow) contains 3×3 deconvolutions and 3×3 convolutions to reduce grid artifacts.

layer (i.e., 112 for 3×3 and 384 for 1×1 convolutions). In the LFN, the information from the two domain is processed into two different streams, obtained by building two towers made of four 3×3 convolutional kernels without sharing the weights between them, in order to learn domain specific features representations. The outputs of the two towers are then concatenated and processed by the final 1×1 convolutions. Final outputs pass through a Sigmoid activation as for CCNN. The number of channels are the same as for EFN model. Both models have been trained and compared with CCNN, proving to perform better when trained with a much larger amount of samples compared to the amount (i.e., 94 stereo pairs versus 20) typically deployed in this field [1]. The receptive field of both networks is the same of CCNN (9×9).

3.2 Proposed global approach

In this section, we describe the network architecture designed to infer confidence prediction by looking at the whole image and disparity content.

ConfNet. Inspired by recent works in stereo matching [28,30,29], we design an encoder/decoder architecture enabling a large receptive field and at the same time maintaining the same input dimensions for the output confidence map. Figure 3 shows an overview of the ConfNet architecture. After concatenating features computed by 3×3 convolutional layers from both RGB reference image and disparity map, they are forwarded to the first part of the network, made of 4 encoding blocks. Each of them is made of a 3×3 convolutional layer ReLU activations and a 2×2 max-pooling used to decimate the input dimension and thus to increase the receptive field. More precisely, after the fourth block the original resolution is reduced by a factor 16, making a 3×3 convolution actually processing a 48×48 receptive field of the initial input. The number of channels of the convolutional layers in different blocks are respectively 64, 128, 256 and

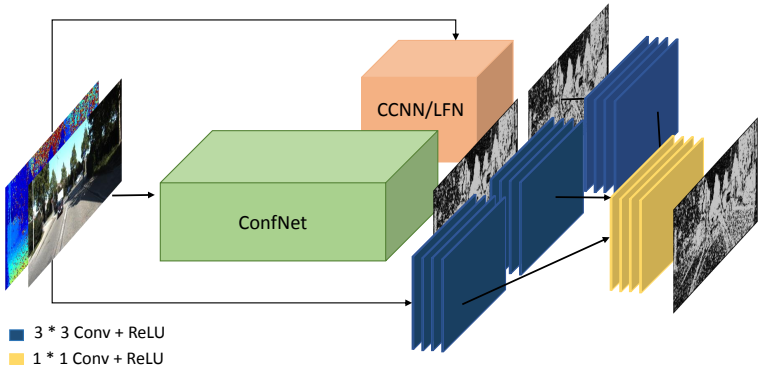

Fig. 4. LGC-net architecture. Given the input reference image and its disparity map, they are forwarded to both local (CCNN or LFN, in orange) and global (ConfNet, green) networks, whose outputs and disparity are processed by 3 independent towers, concatenated to finally infer the output confidence map.

512, doubling after each max-pooling operator. Then, four decoding block follow in order to restore the original resolution of the input before obtaining the final confidence map. Each block uses a 3×3 deconvolutional layer with stride 2, followed by a 3×3 convolutional layer processing deconvolutional outputs concatenated with features taken from the encoding part at the same resolution. This reduces grid artifacts introduced by deconvolutional layers as suggested in [28], as well as enables to keep fine details present before down-sampling in the encoding part and missing after up-sampling from lower resolutions. The number of channels in each block for both deconvolutional and convolutional layers are respectively 256, 128, 64 and 32. A final 3×3 convolutional layer produces the final, full resolution confidence map followed by a Sigmoid operator to obtain normalized confidence values. The much larger receptive field enables to include much more information when computing per-pixel scores, but also acts as a *regularizer* yielding smoother confidence estimations and this leads to poor accuracy when dealing with high frequency patterns.

3.3 Local-global approach

To effectively combine both local and global cues, we introduce a final module acting in cascaded manner after the first two networks by processing their outputs and the initial disparity map. The module in charge of combining these cues is made of three towers processing respectively the local map, the global map and the disparity map. Weights are not shared between towers to extract distinct features from the three domains. Each tower is made of four convolutional layers with kernels 3×3 and 64 channels, their output are then concatenated and forwarded to two final 1×1 convolutional layers producing 100 features map each and a final 1×1 convolution in charge of the final confidence estimation, passed

through a Sigmoid layer. Figure 4 describes the overall framework, referred to as Local Global Confidence Network (LGC-Net).

4 Implementation details and training protocol

We implemented our models using the TensorFlow framework. In particular, we deployed CCNN, EFN and LFN using the same configuration proposed in [2]: 64 and 100 channels respectively for 3×3 and 1×1 convolution, for which we report extensive experimental results in the next section. We also evaluated 112 and 384 channels versions, we report these experiments in the supplementary material. While the entire framework is fully differentiable from the input to the output, thus trainable in end-to-end manner, we first train the local and global networks separately, then we train the cascaded module. As already highlighted in [30], training cascaded models in end-to-end fashion may lead the network to converge at a local minimum, while a reasoned training of each module may enable better overall performance.

Local networks training schedule. Following the guidelines provided in [1], we extract 9×9 image patches from the first 20 stereo pairs in the KITTI 2012 training dataset [9] centered on pixels with available ground-truth disparity used to obtain confidence ground-truths (more details in Section 5.1), resulting into about 2.7 million samples. We trained for 14 epochs as proposed in [2,6] using a batch of dimension 128, resulting into nearly 300k iterations. We used Stocastic Gradient Descent optimizer (SGD) to minimize the Binary Cross Entropy (BCE) [2,6], a learning rate of 0.003 decreased by a factor 10 after 11 epochs and a momentum of 0.9.

ConfNet training schedule. We train ConfNet on 256×512 images estimating a confidence value for each pixel differently from local approaches that estimate confidence only for the central one in a patch (thus requiring to center the neighborhood on a pixel with available ground-truth). Despite training complex architectures like DispNet requires a large amount of data usually obtained from synthetic datasets [28], we found out that training the same 20 images from KITTI is enough to effectively learn a confidence measure. This is probably due to the simpler task the network is faced with. In fact, finding outliers in a disparity map (i.e., a binary classification of the pixels) is much easier compared to infer depth from a stereo pair. Moreover, the disparity domain is less variegated than its RGB counterpart. Despite RGB data being processed jointly with disparity inside ConfNet, it plays a minor role compared to the latter. Cross-validation on Middlebury v3 dataset [10], with indoor imagery extremely different from outdoor environments observed at training time will confirm this fact. We train ConfNet for 1600 epochs extracting random crops from the training stereo pairs, for a total of about 32k iterations. It is worth to note that, at training time, local networks produce a single pixel prediction versus the 256×512 available from ConfNet. For a single iteration, the minimized loss function encodes the contribution from 128 pixels for local networks (i.e., one for each sample in the batch) and 2^{16} for ConfNet, processing $512\times$ the amount of data. For this rea-

sons only 32k iterations are enough for ConfNet to converge compared to the 300k of local methods. Pixels whose disparity ground-truth is not available are masked when computing the loss function. We used SGD and BCE as for local networks, with an initial learning rate of 0.003, divided by a factor 10 after 1k epochs.

LGC-Net final training schedule. Finally, we train the cascaded module after freezing the weights of the local and global networks. We run additional 14 epochs processing image patches extracted from both disparity, local and global confidence estimations. The same 20 images, SGD, BCE loss, learning rate schedule and momentum are used for this training as well.

5 Experimental results

In this section, we report extensive experimental results supporting the superior accuracy achieved by the proposed LGC-Net compared to state-of-the-art. We evaluate the newly proposed framework estimating confidence for disparity maps obtained from three popular algorithms standard in this field [1], respectively AD-CENSUS [11], MC-CNN-fst matching cost [8] and SGM [12]. For this latter algorithm, compared to [1], we tuned better P1 and P2 penalties to 3 and 0.03, obtaining more accurate disparities on KITTI datasets slightly reducing accuracy on Middlebury v3 dataset. In Section 5.1 we outline the evaluation protocol we follow to validate our method, in Section 5.2 we report results on both KITTI 2012 dataset [9] (i.e., on images not involved in training) and KITTI 2015, while in Section 5.3 we cross-validate on Middlebury v3 [10] as commonly done by most recent works [1] to measure how well the confidence measures perform on data quite different from the one deployed for training.

5.1 Evaluation protocol

The standard task on which confidence measures are evaluated is outliers detection [13,1]. It consists in assigning to each disparity assignment a score between 0 an 1 according to their estimated uncertainty. Following the guidelines of standard evaluation benchmarks [9,7,10], each pixel p of an image is considered correctly assigned if its disparity $d(p)$ and its ground-truth label $\tilde{d}(p)$ are distant less than a threshold τ, i.e. $|d(p) - \tilde{d}(p)| < \tau$. The threshold value is assigned according to dataset specifications, in particular for KITTI 2012 and 2015 τ usually it is 3 and for Middlebury v3 it is 1 [1]. The same criterion is used to produce confidence ground-truth labels for training, encoding correct pixels with a score of 1 and outliers with 0. Since in our experiments the training has been always carried out on 20 images of the KITTI 2012 dataset, τ is set to 3 to generate labels. To quantitatively evaluate how well a confidence measure tackles this task, ROC curve analysis represents the standard in this field [13,1]. By plotting the percentage of outliers ε as a function of the amount of pixels sampled from a disparity map in order of decreasing confidence, we can compute the Area Under the Curve (AUC) and average it over the entire evaluation

dataset. The lower is the AUC value, the more accurate is the confidence estimation for outliers detection purpose. The lower bound on a single disparity map is obtained according to its error rate ε as

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

5.2 Evaluation on KITTI datasets

To assess the effectiveness of LGC-Net, we train the networks on the first 20 images of the KITTI 2012 dataset and we report extensive experimental results on the remaining 174 images of the same stereo dataset [9] as well as on the entire KITTI 2015 dataset [7]. This second dataset depicts outdoor environments similar to the first dataset but with the addition of dynamic objects not present in the other. We evaluate confidence measures provided by standalone modules (i.e., CCNN, EFN, LFN and the global architecture ConfNet) as well as those produced by the full local-global framework in two configurations obtained respectively by deploying CCNN [2] or multi-modal architectures [6] as local network. For a fair comparison, all the evaluated models have been trained from scratch following the same protocol described in Section 4. Source code is available here `https://github.com/fabiotosi92/LGC-Tensorflow`.

Table 1 reports experimental results on KITTI 2012. Each row refers to one of the three considered algorithms, respectively AD-CENSUS, MC-CNN and SGM and each column to a confidence measure, reporting AUC values averaged on the entire dataset. In bold, the best AUC for each algorithm. Considering at first single networks, we observe that multi-modal network LFN perform similarly to CCNN being this latter method outperformed by a small margin only with AD-CENSUS. The EFN network has always worse performance compared to CCCN and LFN. These results highlight that, with LFN and EFN networks in this configuration, processing the RGB image does not provide additional information compared to the one inferred from the disparity domain. Looking at ConfNet we can observe how processing global information only leads, as expected, to less accurate results with noisy disparity maps provided by AD-CENSUS but it performs reasonably well, and better than EFN, with smoother disparity maps generated by SGM and MC-CNN. In particular it is always outperformed by CCNN and LFN. According to these results, confirmed also by following evaluations, the global approach alone loses accuracy when dealing with fine details, despite the deployment of skip-connection between encoder and decoder sections, while local approaches performs very well in these cases. Observing LGC-Net results, both configurations outperform all the other evaluated techniques, highlighting how the two complementary cues from local and global networks can be effectively combined to improve confidence estimation moving a step forward optimality for all the three stereo algorithms. By directly comparing the two configurations of LGC-Net, using respectively CCNN or LFN as local network, there is no clear winner highlighting how the contribution given by the RGB image on a small neighborhood seems negligible. In fact, it yields

KITTI 2012 [9] (174 images)	CCNN [2]	EFN [6]	LFN [6]	ConfNet	LGC-Net (CCNN)	LGC-Net (LFN)	Optim.
AD-CENSUS [11]	0.1207	0.1261	0.1201	0.1295	**0.1174**	0.1176	0.1067
MC-CNN [8]	0.0291	0.0316	0.0294	0.0311	0.0279	**0.0278**	0.0231
SGM [12]	0.0194	0.0229	0.0198	0.0199	0.0176	**0.0175**	0.0088

Table 1. Experimental results on KITTI 2012 dataset [9]. From top to bottom, evaluation concerning AD-CENSUS [11], MC-CNN [8] and SGM [12] algorithms. For each column, average AUC achieved on the entire dataset (i.e., 174 out of 194 stereo pairs) for different confidence measures.

KITTI 2015 [7] (200 images)	CCNN [2]	EFN [6]	LFN [6]	ConfNet	LGC-Net (CCNN)	LGC-Net (LFN)	Optim.
AD-CENSUS [11]	0.1045	0.1087	0.1026	0.1128	**0.0999**	0.1004	0.0883
MC-CNN [8]	0.0289	0.0319	0.0292	0.0315	0.0281	**0.0278**	0.0213
SGM [12]	0.0201	0.0239	0.0209	0.0216	0.0193	**0.0190**	0.0091

Table 2. Experimental results on KITTI 2015 dataset [7]. From top to bottom, evaluation concerning AD-CENSUS [11], MC-CNN [8] and SGM [12] algorithms. For each column, average AUC achieved on the entire dataset (i.e., 200 stereo pairs) for different confidence measures.

a 0.0001 difference in terms of average AUC between the two versions, in favor of the first configuration on AD-CENSUS and the second one on MC-CNN and SGM. These experiments highlights that the major benefit is obtained by the proposed strategy exploiting local and global context information.

Table 2 reports experimental results on the KITTI 2015 dataset [7], with AUC values averaged over the available 200 stereo pairs with ground-truth. First of all, we observe that the same trend observed for KITTI 2012 is confirmed also in this case, with CCNN being slightly outperformed by LFN only on AD-CENSUS. CCNN and LFN always provide more accurate estimation accuracy compared to EFN while ConfNet outperforms this latter method on smoother MC-CNN and SGM disparity maps as in previous experiment. Finally, the two LGC-Net versions achieve overall best performance on this dataset, as for KITTI 2012, confirming the effectiveness of the proposed method. Moreover, the same results also highlight once again the negligible margin brought in by using the RGB image with CCNN.

5.3 Cross-validation on Middlebury v3

Having proved the effectiveness of the proposed LGC-Net on KITTI datasets, we conduct a more challenging evaluation by cross-validating on Middlebury v3 imagery [10] confidence measures trained on the first 20 images of the KITTI 2012 dataset. As done in [1], assessing the performance on a validation dataset quite different from the one used during the training phase effectively measures how robust a confidence measure is with respect to circumstances very likely to occur in practical applications. Being our models trained on KITTI images, de-

Middlebury v3 [10] (15 images)	CCNN [2]	EFN [6]	LFN [6]	ConfNet	LGC-Net (CCNN)	LGC-Net (LFN)	Optim.
AD-CENSUS [11]	0.1131	0.1263	0.1146	0.1206	**0.1099**	0.1109	0.0899
MC-CNN [8]	0.0668	0.0781	0.0645	0.0755	0.0624	**0.0616**	0.0458
SGM [12]	0.0794	0.1005	0.0856	0.0886	**0.0703**	0.0709	0.0431

Table 3. Experimental results on Middlebury v3 dataset [10]. From top to bottom, evaluation concerning AD-CENSUS [11], MC-CNN [8] and SGM [12] algorithms. For each column, average AUC achieved on the entire dataset (i.e., 15 stereo pairs) for different confidence measures.

picting outdoor environments concerned with autonomous driving applications, the indoor scenes included in the Middlebury v3 dataset represent a completely different scenario ideal for the kind of cross-validation outlined.

Table 3 quantitatively summarizes the outcome of this evaluation. First and foremost, as in previous experiments, LGC-Net outperforms with both configurations all standalone confidence measures confirming the negligible difference, lower or equal than 0.001, between the two local networks. The trend between single architectures is substantially confirmed with respect to previous experiments, with ConfNet performing always better than EFN in this cross-evaluation even with the noisy AD-CENSUS maps. CCNN and LFF, as for previous experiments, performs quite similarly confirming once again the small impact of RGB cues in local networks with our training configuration.

In Figure 5 we report a qualitative comparison between local, global (ConfNet) and LGC-Net for two images of the the Middlebury v3 dataset processed with SGM and MC-CNN stereo algorithms. The quantitative advantages reported for LGC-Net in the previous evaluations can be clearly perceived qualitatively by looking, for instance, at texture-less regions on the wall in *PianoL* stereo pair and at the occluded area on the background in *Pipes* stereo pair.

To summarize, exhaustive experiments on three datasets and three stereo algorithms proved that the proposed framework always outperforms both local and global standalone strategy by a significant margin, thus effectively learning to combine local and global cues to obtain more accurate confidence estimation. This trend is also confirmed moving to very different data as reported in the cross evaluation, proving that LGC-Net is more capable to generalize to completely different image contents. Overall, the proposed method always outperforms state-of-the-art methods for confidence estimation.

6 Conclusions

In this paper we propose, for the first time to the best of our knowledge, to leverage on global and local context to infer a confidence measure for stereo. Driven by the outstanding results achieved by CNN-based confidence measures, in this paper we argue that their effectiveness can be improved by changing their intrinsic local nature. To this aim we propose to combine with a CNN cues inferred with two complementary strategies, based on two very different receptive

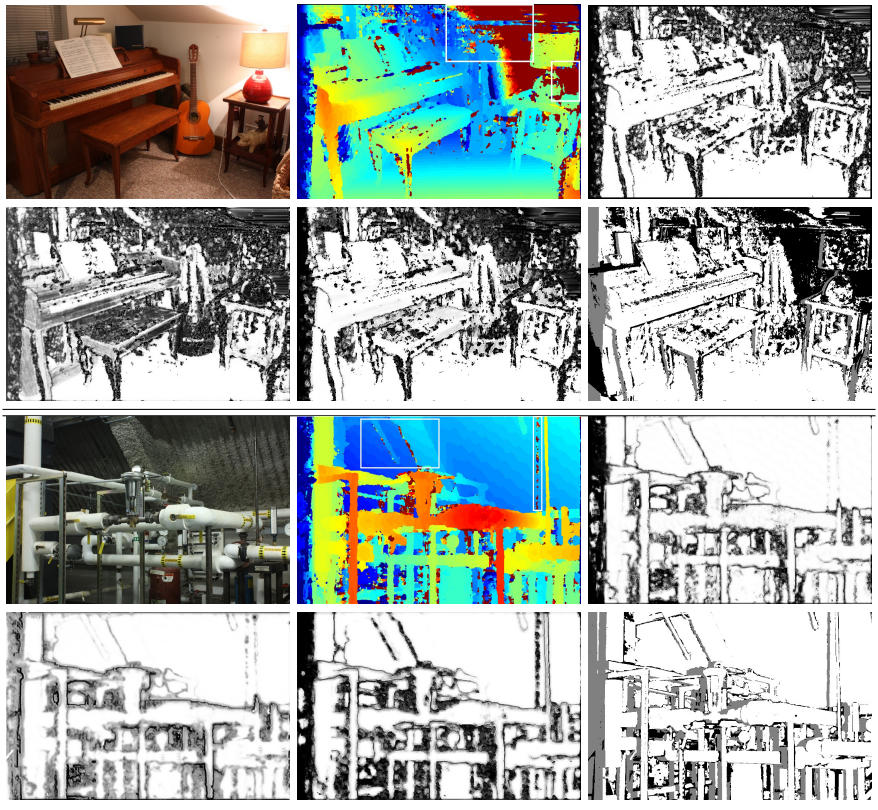

Fig. 5. Qualitative comparison of confidence maps on selected images from Middlebury v3 dataset [10]. For each sample, we report from top left to bottom right reference image, disparity map, confidence map respectively for CCNN, ConfNet and LGC-net and ground-truth confidence labels. On top *PianoL* pair processed by MC-CNN-fst, on bottom *Pipes* pair processed by SGM.

fields. The proposed LGC-Net, a multi-modal cascaded network, merges the outcome of the two complementary approaches enabling more accurate confidence estimation. We extensively evaluated the proposed method on three datasets and three algorithms following standard protocols in this field proving that our proposal outperforms state-of-the-art confidence measures and further moves a step forward optimality.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We also thank Alessandro Fusco for his preliminar experiments on the ConfNet architecture.

References

1. Poggi, M., Tosi, F., Mattoccia, S.: Quantitative evaluation of confidence measures in a machine learning world. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
2. Poggi, M., Mattoccia, S.: Learning from scratch a confidence measure. In: Proceedings of the 27th British Conference on Machine Vision, BMVC. (2016)
3. Seki, A., Pollefeys, M.: Patch based confidence prediction for dense disparity map. In: British Machine Vision Conference (BMVC). (2016)
4. Poggi, M., Mattoccia, S.: 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)
5. Haeusler, R., Nair, R., Kondermann, D.: Ensemble learning for confidence measures in stereo vision. In: CVPR. Proceedings. (2013) 305–312 1.
6. Fu, Z., Ardabilian, M.: Stereo matching confidence learning based on multi-modal convolution neural networks. In: Representation, analysis and recognition of shape and motion FroM Image data (RFMI). (2017)
7. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
8. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research **17**(1-32) (2016) 2
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3354–3361
10. Scharstein, D., Hirschmller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In Jiang, X., Hornegger, J., Koch, R., eds.: GCPR. Volume 8753 of Lecture Notes in Computer Science., Springer (2014) 31–42
11. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Proceedings of the Third European Conference on Computer Vision (Vol. II). ECCV '94, Secaucus, NJ, USA, Springer-Verlag New York, Inc. (1994) 151–158
12. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 807–814
13. Hu, X., Mordohai, P.: A quantitative evaluation of confidence measures for stereo vision. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2012) 2121–2133
14. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision **47**(1-3) (apr 2002) 7–42
15. Spyropoulos, A., Komodakis, N., Mordohai, P.: Learning to detect ground control points for improving the accuracy of stereo matching. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2014) 1621–1628
16. Park, M.G., Yoon, K.J.: Leveraging stereo matching with learning-based confidence measures. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015)
17. Poggi, M., Mattoccia, S.: Learning to predict stereo reliability enforcing local consistency of confidence maps. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)

18. Poggi, M., Tosi, F., Mattoccia, S.: 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)
19. Kim, S., Min, D., Kim, S., Sohn, K.: Feature augmentation for learning confidence measure in stereo matching. IEEE Transactions on Image Processing **26**(12) (2017) 6019–6033
20. Mostegel, C., Rumpler, M., Fraundorfer, F., Bischof, H.: Using self-contradiction to learn confidence measures in stereo vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. (2016) 4067–4076
21. Tosi, F., Poggi, M., Tonioni, A., Di Stefano, L., Mattoccia, S.: Learning confidence measures in the wild. In: 28th British Machine Vision Conference (BMVC 2017). (September 2017)
22. Poggi, M., Tosi, F., Mattoccia, S.: Efficient confidence measures for embedded stereo. In: 19th International Conference on Image Analysis and Processing (ICIAP 2017). (September 2017)
23. Marin, G., Zanuttigh, P., Mattoccia, S.: Reliable fusion of tof and stereo depth driven by confidence measures. In: 14th European Conference on Computer Vision (ECCV 2016). (2016) 386–401
24. Poggi, M., Mattoccia, S.: Deep stereo fusion: combining multiple disparity hypotheses with deep-learning. In: Proceedings of the 4th International Conference on 3D Vision, 3DV. (2016)
25. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
26. Tonioni, A., Poggi, M., Mattoccia, S., Di Stefano, L.: Unsupervised adaptation for deep stereo. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
27. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1592–1599
28. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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). (June 2016)
29. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
30. Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)