Multi-level and Multi-modal Feature Aggregation for Stereo Matching Confidence Estimation

simonren@sjtu.edu.cn

Abstract. Confidence estimation for stereo matching is an essential task since it can be used to further improve stereo matching in the post-processing step. Moreover, it provides the reliability information to help automated driving systems to measure distance just by relying on high confidence. In this paper, we propose a convolutional neural network (CNN) based architecture that is able to extract multi-level cues from low-level to high-level feature extractor layers, in the RGB and disparity domains, and to elaborate on them to obtain a more accurate and robust confidence estimation. In order to achieve this, we train an end-to-end CNN model which deploys a multi-level feature extractor layers with different dilation to obtain the different scales of the receptive fields. In this way, the feature maps extracted by the deeper layers of a convolutional network encode higher-level semantic information and multi-scale context information contained in the large receptive field of each neuron, while the shallower layers encode more local information. By concatenating multi-level feature cues from the disparity map and its reference image, we can aggregate multiple features from different scales incorporating patch-based features with multi-scale region-level context information to complement the power of different feature layers and different domain inputs. In this paper, we highlight that it is indispensable to fuse multi-level features from not only disparity map but also its reference image to predict accurate and robust confidence. The experimental results on three well-known datasets for automated driving as well as with two popular stereo algorithms clearly show that the proposed approach outperforms state-of-the-art confidence estimation methods.

1 Introduction

Stereo vision plays an important and fundamental role to infer the 3D structure of real-world imagery and for this reason deployed in several computer vision applications such as autonomous vehicles [1], robotic navigation [2], object detection and recognition [3]. Typically for two images of different views on the same scene, taken by cameras with horizontal displacements, the task of stereo matching is to find the corresponding pixels between the left and right images to infer depth through simple triangulation. The displacement between the corresponding pixels is called disparity and the set of all disparities in the image

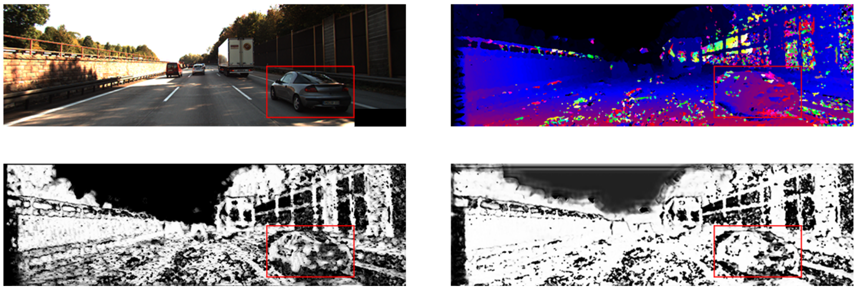


Fig. 1. From top left to bottom right, it is the left image of frame 196 from KITTI 2015, disparity map calculated used our modified SGM version named census-SGM, confidence estimation results estimated by LGC-CCNN and our network MMFA-LFN respectively. We highlight a car for each image in the same position.

is called the disparity map. Stereo vision, despite being one of the most active areas of research, still suffers from various issues, such as occlusions [4], transparent or reflective surfaces, and texture-less or repeated pattern regions [5, 6], which are prone to find out mismatch disparities. These limitations lead to incorrect disparity assignments, which limits the adaptability of stereo disparity estimation for practical vision applications. Over the years, different approaches have been proposed to improve dense disparity estimation. The stereo matching pipeline is classically decomposed into three or four steps: matching cost computation, matching cost aggregation [7], disparity prediction and optional disparity refinement [8]. Nowadays, the post-processing step for the refinement of the disparities is taken into account and widely applied to improve the robust and accurate result of disparity estimation in the almost all stereo matching algorithms. Confidence measures (CMs) have been proposed to rate the correctness of matches, which can be applied to improve stereo matching in the post-processing step. Generally, the very first step in this process is to filter out the wrong assignments by means of effective confidence measures (CMs) aimed at encoding the degree of uncertainty of each pixel and then correct them using reliable pixels [9–11].

Based on the used data cues, conventional confidence measures can be categorized into three groups:

1. **Confidence measures employing matching cost volumes:** CMs in this category are related to the minimum or the second minimum matching cost, or a combination of both. In [12], we can see that the classic examples of these approaches are Naive Peak Ratio (PKRN) , Maximum Likehood Measure (MLM) and Left-Right Difference (LRD) .
2. **Confidence measures utilizing initial disparity maps:** Initial disparity maps are used as input cues to decode the CMs. Examples of these CMs approaches are Left-Right Consistency (LRC) [13], Variance of the Disparity Values and the Median Deviation of Disparity Values (MDD) [14].

3. **Confidence measures based on source images pairs:** This category of confidence measure includes Magnitude of the Image Gradients Measure (MIGM) [9] and Distance to Border (DB) [14].

Such hand-crafted features used in conventional confidence measures are dependent upon expertise and knowledge in stereo vision to perform well on detecting the mismatches for some certain challenges. In order to address this weakness, many works [14, 9, 15] have been proposed that a machine learning framework based on the random forest (RF) can achieve the optimal results when jointly processing a pool of CMs. For instance, Park & Yoon in [15] proposed to extract different cues from the estimated disparity map and cost volume, which were then used to train a simple classifier, a random forest to predict the confidence of correct correspondence.

Although the combined features which are thoughtfully formulated and selected to well train a random forest framework, it cannot guarantee that all discriminating information has been incorporated. In addition, such methods still rely on hand-crafted features. Recently, researchers demonstrated that CNN learning-based confidence measures [16–18] can achieve a state-of-the-art performance without relying on any hand-crafted features. Generally, cnn-based confidence measures which use information extracted from the disparity map [16] or both from the disparity map and reference image simultaneously [17, 18] as input cues. CNN-based methods have the good capability of adapting to different data.

In this paper, we present a novel and high-performance confidence estimation network, named multi-level and multi-modal feature aggregation network (MMFA-net), for accurate confidence estimation from two modalities, the disparity map and its reference image. CNN features of different layers with increased dilation factor enlarging receptive field aim to encode different-level information. Multiple features from high-layers care more about multi-scale context information, while features from low-layers contain more local information. Our MMFA-network fuses multifarious features both from the disparity map and its reference image to exploit complementary strengths and different cues from different input domains. The contributions of this paper are as follows:

1. We propose an end-to-end learning framework for confidence estimation without separately training different models.
2. In this paper, we highlight that it is indispensable to fuse multi-level features from not only the disparity map but also its reference image simultaneously to predict accurate and robust confidence. We propose network architectures which can effectively aggregate multiple features from different scales incorporating patch-based features with multi-scale region-level context information to complement the power of different feature layers and different domain inputs.
3. we extensively evaluated the performance of the proposed framework on three popular datasets for automated driving, KITTI2012 [19], KITTI2015 [20], and *Driving* [21] using two different stereo algorithms SGM [7] and

ADCensus [6]. Experimental results prove that our proposal can achieve a state-of-the-art performance.

2 Related Work

2.1 Hand-Crafted Methods

In literature, many methods have been proposed to estimate the reliability of disparities by relying on manually designed features [22, 23]. Based on the evaluation results in [12], a single feature is not sufficient to estimate the accurate and robust confidence. To alleviate the weakness of separate measures, many works [15, 14, 24, 9] proposed to combine various features. Generally, features are extracted from the estimated disparity map and cost volume as input cues to train a simple classifier, for example, a random forest. In particular, Park and Yoon [15] first analyzed the specialty of various confidence measures and selected the optimal confidence features from multiple confidence features using a regression forest. Then with the feature vectors of selected measures, they trained another random forest to further improve the confidence estimation. More recently, Poggi and Mattoccia [16] took advantage of confidence features extracted from the disparity map to train an ensemble classifier similar to [15, 14, 24, 9] while achieving better results within a time complexity of O(1). In [25], they extracted cues from the cost volume to predict pixel-wise reliability of the estimated disparity and achieved much better results. In [26], spatial context was introduced to estimate confidences at the superpixel-level compared to all the above explained methods working at the pixel-level.

2.2 CNN-Based Methods

CNNs have been successfully applied to many computer vision applications. Due to their outstanding feature learning capabilities, many researchers proposed CNN-based methods to estimate dense disparity [27–29] and boost the performance compared to the traditional methods. Recently, outstanding works also [16–18] have been proposed to apply CNNs to estimate confidence. Poggi and Mattoccia [16] extracted a square patch centered on the disparity map and forwarded it to a CNN, trained to distinguish between patterns corresponding to correct and erroneous disparity assignments and, thus, to infer a confidence value. In [17], Fu and Ardabilian proposed a multi-modal deep learning approach for stereo matching confidence estimation. The input of their method was comprised of two modalities, the initial disparity map, and its reference color image. Due to the limitations of small receptive fields in [16, 17], furthermore, in [18] they deployed 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. Specifically, by training a multi-modal cascaded architecture they first obtained two confidence predictions by reasoning respectively on local and farther cues, then they further elaborated on them to obtain a final, more accurate prediction. Instead of using the disparity map and reference image as inputs, Kim *et al.* [30] proposed a convolutional neural network to predict disparity and confidence simultaneously by taking a row cost volum as inputs. Poggi *et al.* extensively evaluated state-of the-art stereo confidence estimation methods and showed that deep learning-based approaches have significant advantages in efficiency and accuracy [12].

3 Proposed Method

We first explain as a background how our proposed method is motivated based on our careful and insightful observations in Section 3.1. In Section 3.2, we describe the proposed network architecture named MMFA-net. Subsequently, the strategy for training the network is discussed in Section 3.3.

3.1 Background

Our method is motivated and driven by the following observations. These observations should be carefully considered to design a effective and efficient CNN architecture to boost performance.

In [16], it has been shown that local feature from disparity map can clearly assess the reliability of the disparity assignments since local regions in the disparity map often contain recurrent patterns which characterize correct and incorrect disparity assignments. However, local feature suffers from the problem of semantic ambiguity and context information. Observing that, to some extent, disparity map calculated from a pair of rectified stereo images under some specific viewpoints can provide semantic information and thus extracting context information from the disparity map becomes significant. As illustrated in Fig.1, we can directly distinguish the car from the disparity map. Intuitively, if we can perceive each object in the disparity map, we can basically deduce the distribution changes of disparities for each object. Thus, such cues should be fully used to complement the limits of local feature.

It is obvious that we can recognize some objects in the disparity map at first glance. However, it is still difficult to distinguish the objects from the disparity map because the mismatch disparities are calculated in inherently ill-posed regions which include occlusion areas, repeated patterns, texture-less regions, and reflective surfaces. Inversely, it is easier to extract semantic information from the reference image since we can label each object (*e.g.* trees, street, sky, and driving cars) correctly in the reference image. In other words, the reference image can help to identify the objects in the disparity map. So it is necessary to combine with the high-level features encoding context information from the reference image to capture the objects for confidence estimation. Capturing the objects from the disparity map and reference image aims to help confidence estimation. However, a challenge is caused by the existence of objects at multiple scales (*e.g.* driving cars in different distances from the view). Thus multi-level features encoding multi-scale context information should be incorporated into

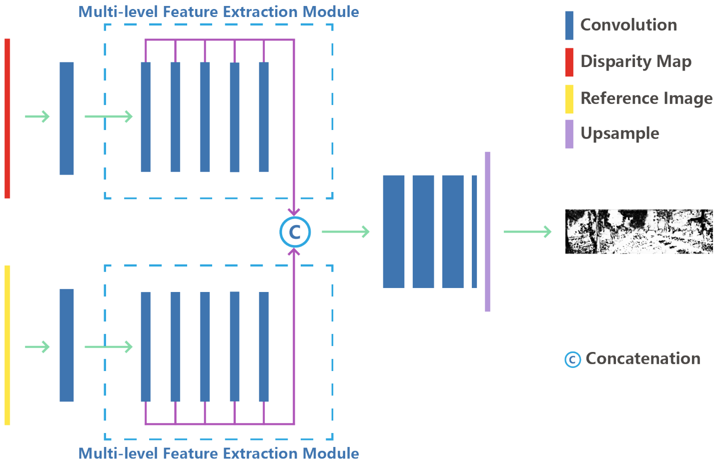

Multi-level Feature Extraction Module
Convolution
Disparity Map
Reference Image
Upsample
C
Concatenation
Multi-level Feature Extraction Module

Fig. 2. Proposed network architecture (MMLF-Net)

the confidence estimation to successfully recognize both large and small objects from the disparity map and its reference image.

Also, it is reasonable that the patch-level features extracted from the reference image plays a vital role in predicting confidence. A square patch centered on the reference image can reflect the detailed structure information compared with the context information. For example, photometric deformations, bad illumination, and transparent texture may lead to wrong disparity assignments with high probability, in other words, it indicates a low confidence scores should be rated to the corresponding disparities in such local regions.

3.2 Multi-level and Multi-modal Feature Aggregation Network

Based on the analysis in Section 3.1, local features as well as multi-scale context information both from the disparity map and its reference image are should be fused to improve the performance of confidence estimation. In this section, we present our MMFA neatly to achieve these goals.

In a CNN, the dilated convolution operator can apply the same filter at different ranges using different dilation factors. The dilation factors increase as the layers go deeper to increase the receptive field and capture contextual information without losing the spatial resolution compared to those from encoder-decoder architectures. Thus, we increase dilated factors as the layers go deeper to encode different-level features at different layers. In MMFA-net, high-layer features care more about semantic information and context information, while low-layer features focus more on detailed information. The input of our method is comprised of two modalities, the initial disparity map and its reference image. We propose to exploit complementary strengths of different layers by fusing multi-level features of the reference image and disparity map. In this way, multi-level features are in charge of processing the local features and multi-scale context respectively, while multi-modal features help to overcome the limitations by just taking the disparity domain into account. It is significantly different from previous local patch-based CNN architectures [16]. We not only explicitly consider the path-based features but also the multi-scale context information extracted from disparity maps. In contrast to the architecture in [17], their method lacks context information to predict confidence by only taking advantage of patch-based features from two modalities, the disparity map and its reference image. Although the method in [18], it exploits more global context both from disparity maps and reference image by enlarging receptive field with encoder-decoder architecture for learning confidence predictions, it doesn't consider multi-level features encoding multi-scale cues including patch-based features and multi-scale context information which can provide more multiple information to help confidence estimation. In addition, the model proposed in [18] needs to be separately trained with more parameters compared with our network architecture.

There are different ways to fuse multi-level features from the disparity map and its reference image. Similar to the work in [17], we propose two versions of MMFA-net according to their strategy of fusing.

Early Fusion Network: Disparity map and its reference image are concatenated to form a 4-channel image, which is then fed into the network that only has one branch to extract multi-level features. The network architecture consists of a series of convolutions layers. Specifically, the input is first processed by a 3×3 convolutional layer with stride 2, and then followed by the multi-level feature extraction module. Multi-level feature extraction module has 5 layers that apply 3×3 convolutions with increased dilation factors at different scales. The dilations are 1, 1, 2, 4, and 8 respectively. Note that dilated convolutions support exponential expansion of the receptive field and thus extract multi-level features effectively. Subsequently, the outputs of different layers in multi-level feature extraction module are then concatenated and processed by a 1×1 convolutional layer. This is followed by two 3×3 convolutional layers to further refine the fused features. A final 3×3 convolutional layer produces the final confidence map followed by a Sigmoid operator to obtain normalized confidence values, after which the low-dimensional confidence map is upsampled to the same size of the original disparity map via bilinear interpolation. The number of convolutional filters in each convolution layer from bottom to top are 64, 32, 32, 32, 32, 32, 256, 256, 256, and 1 respectively. All convolutional layers are followed by BN and ReLU except the last one.

Late Fusion Network: As shown in Fig.2, in contrast to the early fusion version, it contains two dependent branches separately extracting multi-level features from the disparity map and reference image respectively without sharing weight as siamese networks. Apart from this, all network architecture settings are the same to the early fusion version. In order to aggregate multi-level and multi-modal features, we concatenate the features extracted from different-level convolutional layers in two multi-level feature extraction modules respectively.

Table 1. Experimental results on KITTI2012 and KITTI2015 datasets. For each row, average AUC achieved on the entire dataset (174 out of 194 stereo pairs for KITTI2012 and 200 stereo pairs of KITTI2015) is listed for different confidence network architectures. The disparity map is calculated with census-SGM, and ADCensus respectively.

DataSet	KITTI 2012		KITTI 2015	
Stereo Algorithm	census-SGM	ADCensus	census-SGM	ADCensus
CCNN [16]	0.06239	0.13032	0.06018	0.10937
EFN [17]	0.06621	0.13198	0.06358	0.12233
LFN [17]	0.06278	0.09903	0.06071	0.08682
ConfNet [18]	0.06659	0.09346	0.06823	0.09207
LGC-Net(CCNN) [18]	0.05850	0.08685	0.05659	0.07842
LGC-Net(LFN) [18]	0.05893	0.08808	0.05739	0.07707
MMFA-EFN	0.05524	0.08477	0.05405	0.07021
MMFA-LFN	**0.05488**	**0.07551**	**0.05395**	**0.06263**
Optimal	0.04616	0.04225	0.04248	0.03074

3.3 Training Procedure

The proposed network was implemented using the Pytorch framework. We trained the proposed CNN architectures on the first 20 frames of the KITTI2012 dataset [19] and their corresponding disparity maps calculated by different stereo matching algorithms. During training, images and disparity maps were randomly cropped to size H=256 and W=512. All models were end-to-end trained using an Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with Binary Cross Entropy (BCE) between $\hat{o}$ and ground truth o on each sample i of the mini-batch (1). The $\hat{o}$ is the output of applying a sigmoid function $S\left(x\right)$ (2) to the output of the network.

$$BCE\left(o,\hat{o}\right) = -\tfrac{1}{N}\sum_{i=1}^{N}\left(o_i \cdot \log\left(\hat{o}_i\right) + \left(1-o_i\right) \cdot \log\left(1-\hat{o}_i\right)\right) \tag{1}$$

$$S\left(x\right) = \tfrac{1}{1+e^{-x}} \tag{2}$$

where o_i is the ground truth of the i-th training sample, after calculating the absolute difference between output of different stereo matching algorithms and ground truth disparity map, o_i will be labeled 1 when the absolute difference is less than the given threshold (based on the KITTI benchmark instructions, threshold is 3 here) and 0 for otherwise. The batch size was set to 1 for the training. All networks were trained for 130 epochs with the learning rate set to 0.003 at the beginning and decreased to 0.0003 at the 100-th epoch.

4 Experiment Results

In this section, we report extensive experimental results supporting the superior accuracy achieved by the proposed MMFA-Net compared to state-of-the-art methods, including the learning-based confidence measures such as CCNN [16], multi-modal networks (LFN and EFN) [17], and local-global confidence

Table 2. Experimental results on the *Driving* dataset. For each row, average AUC achieved on the entire dataset (4400 frames of stereo pairs) is listed for different confidence network architecture. The disparity map is calculated with census-SGM, and ADCensus respectively. We also report trainable parameters of different networks.

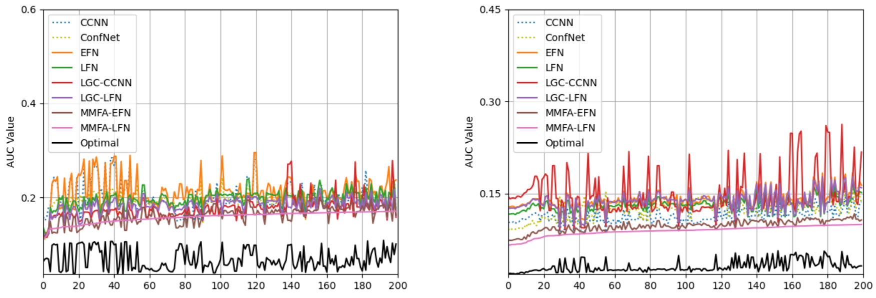
DataSet	*Driving* Scene		Network Parameters
Stereo Algorithm	census-SGM	ADCensus	
CCNN [16]	0.26018	0.32067	128125
EFN [17]	0.29100	0.35781	129853
LFN [17]	0.27777	0.32826	247101
ConfNet [18]	0.27800	0.33471	7855937
LGC-Net(CCNN) [18]	0.30571	0.33298	8347835
LGC-Net(LFN) [18]	0.28529	0.32295	**8466811**
MMFA-EFN	0.24631	0.30819	1280450
MMFA-LFN	**0.23643**	**0.30229**	1377154
Optimal	0.12203	0.14425	-

network (LGC-Net) [18]. We used three famous datasets for automated driving (KITTI2012 [19], KITTI2015 [20], and *Driving* in Scene Flow [21]) and two popular stereo matching algorithms standard in this field, respectively AD-CENSUS [6] and SGM [7] to evaluate the performance of our proposed network. For the SGM algorithm, we implemented a simplified version compared to [7]. Specifically, we set P1 and P2 penalties to 3 and 30 and use census-block matching instead of mutual information to compute initial matching Cost Volume (CV) for which can be deployed in parallel. To compute the CV, we first compute the point-wise matching costs according to the Hamming distance on census transformed images computed on 5×5 patches and then aggregate point-wise matching costs on 5×5 patches for final point-wise matching costs. Next, we employ a multi-direction (8 paths in our implemented version) scanline optimizer to refine the initial CV. The disparity map is obtained from the CV by means of the Winner-Takes-All (WTA) strategy and then the final disparity results are obtained by smoothing the interpolated previous disparity map with a 3×3 median filter for two iterations.

In Section 4.1, we outline the evaluation protocol we follow to validate our method. In Section 4.2 and 4.3, we report the quantitative and qualitative results for all three datasets, respectively, showing the better accuracy achieved by our proposed network.

4.1 Evaluation Methodology

The sparsification curve and its area under the curve (AUC) is a well-known evaluation strategy used to benchmark the capability of the different confidence measures to distinguish correct disparity assignments from erroneous ones [12, 31]. For a confidence map of a given method, all pixels with ground truth are first sorted according to descending confidence. The ordered pixels are then divided into m equal samples (*e.g.* m = 20 in our paper) and each time we pick one



(a)

(b)

Fig. 3. AUC values of (a) ADcensus and (b) census-based SGM for the *Driving* dataset [33]. We sort the AUC values in the ascending order according to proposed MMFA-LFN's AUC values for the first 200 frames. The 'optimal' AUC values are calculated using ground truth confidence map.

sample and calculate its bad pixel error rate (bad pixel defined as differences larger than±3). In this way, we plot the sparsification curves. The Area Under the Curve (AUC) is then used to benchmark the quantitative accuracy of the method. The optimal AUC can be easily deduced as the following formula (3).

$$AUC_{opt} = \int_{1-\varepsilon}^{\varepsilon} \frac{p-(1-\varepsilon)}{p} dp = \varepsilon + (1-\varepsilon) \cdot \ln(1-\varepsilon) \tag{3}$$

Where ε represents the bad pixel rate and p represents pixel density sampled from the disparity map according to descending order of the confidence. Obviously, AUC closer to the optimal value indicates a better confidence prediction.

4.2 Evaluation on KITTI2012 and KITTI2015

To evaluate the performance of the proposed method, we have compared it to state-of-the-art deep learning methods in the literature. As our goal aims to let automated driving systems measure distance relying on the high confidence disparities since the low confidence disparities more likely lead to erroneous results, we choose two famous datasets KITTI2012 and KITTI2015 for automated driving to evaluate our proposed method. KITTI2015 depicts outdoor environments similar to the KITTI2012 but with the addition of dynamic objects not present in the KITTI2012. Thus, it is suitable for assessing the generalization or adaptability of our proposed network as well as state-of-the-art methods. For a fair comparison, all the evaluated models have been trained from scratch following the same protocol described in proposed methods. All models are trained on the first 20 images of the KITTI2012 dataset. Then we report results on the remaining 174 images of the KITTI2012 and on the entire KITTI2015 dataset as shown in Table 1. Our proposed method MMFN-LFN yields the best results.

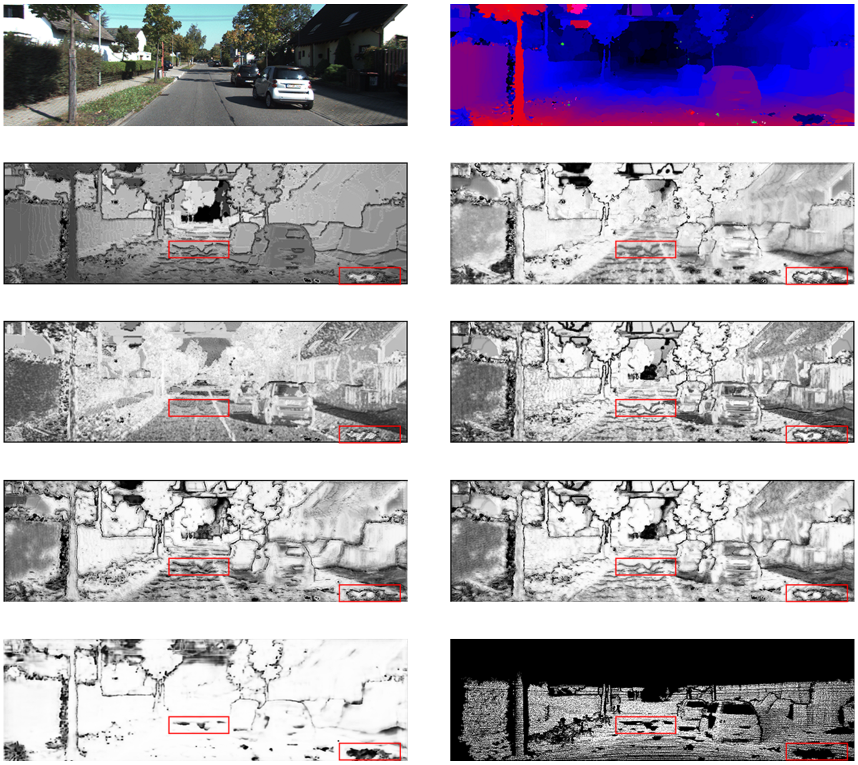


Fig. 4. Qualitative comparison of confidence maps on selected images from KITTI2012 datasets. From top left to bottom right: reference image, disparity map, confidence map respectively for CCNN, ConfNet, EFN, LFN, LGC-CCNN, LGC-LFN, and our proposed MMFA-LFN and ground-truth confidence labels.

We also show qualitative results in Fig.4. It is evident from the qualitative result shown in Fig.4 that our proposed network is more accurate in detecting certain regions.

4.3 Cross-validation on *Driving* dataset

Since our proposed network architecture aims to help the automated driving systems to make a good decision just relying on high confidence, we care more about performance on different outdoor scenes for automated driving. Thus, we have cross-validated our network on the artificial dataset *Driving* scene [21] in order to further demonstrate the ability to generalize when performing on a dataset which is much different from the training dataset. The *Driving* dataset consists of 4400 frames created by Blender (http://www.blender.org) using 3D CAD model with dense ground truth disparity maps. The KITTI datasets are collected from a test vehicle with respect to circumstances occurring in practical autonomous driving applications. As shown in Table 2, our methods can achieve the best results and show the ability of adapting to multiple driving scene, while requiring fewer training parameters compared with the method obtaining the second-best results. From Fig.3, we also can see that our methods, MMFA-EFN and MMFA-LFN have lower overall AUC values.

5 Conclusions

In this paper, we observed that cues from the disparity map and its reference image from multiple scales contribute to accurate confidence estimation and we proposed a novel confidence estimation network (MMFA-network) for confidence estimation. By concatenating different-level features extracting both from the disparity map and its reference image, we can aggregate local features and multi-scale context information to exploit complementary strengths. The experimental results demonstrate that our network architectures achieve considerable better performance in comparison with several state-of-the-art methods. To the best of our knowledge, this is the first method that exploits local feature and multi-scale context from two modalities, the disparity maps and its reference image, to estimate confidence.

References

1. Sivaraman, S., Trivedi, M.M.: A review of recent developments in vision-based vehicle detection. In: 2013 IEEE Intelligent Vehicles Symposium (IV), IEEE (2013) 310–315
2. Schmid, K., Tomic, T., Ruess, F., Hirschmüller, H., Suppa, M.: Stereo vision based indoor/outdoor navigation for flying robots. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE (2013) 3955–3962
3. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems. (2015) 424–432
4. Wildes, R.P.: Detecting binocular half-occlusions: empirical comparisons of five approaches. IEEE Transactions on Pattern Analysis Machine Intelligence (**24**) p.1127–1133
5. Humenberger, M., Zinner, C., Weber, M., Kubinger, W., Vincze, M.: A fast stereo matching algorithm suitable for embedded real-time systems. Computer Vision Image Understanding **114** (2010) 1180–1202
6. Mei, X., Sun, X., Zhou, M., Jiao, S., Wang, H., Zhang, X.: On building an accurate stereo matching system on graphics hardware. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). (2011) 467–474
7. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008) 328–341
8. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47** (2002) 7–42

9. Spyropoulos, A., Komodakis, N., Mordohai, P.: Learning to detect ground control points for improving the accuracy of stereo matching. In: CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1621–1628
10. Hosni, A., Rhemann, C., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. IEEE Trans Pattern Anal Mach Intell **35** (2013) 504–511
11. Xu, L., Jia, J.: Stereo matching: An outlier confidence approach. In: European Conference on Computer Vision. (2008) 775–787
12. Hu, X., Member, S.: P.: A quantitative evaluation of confidence measures for stereo vision. PAMI (2012) 2121–2133
13. Egnal, G., Mintz, M., Wildes, R.P.: A stereo confidence metric using single view imagery. In: PROC. VISION INTERFACE. (2002) 162–170
14. Haeusler, R., Nair, R., Kondermann, D.: Ensemble learning for confidence measures in stereo vision. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. (2013) 305–312
15. Park, M.G., Yoon, K.J.: Leveraging stereo matching with learning-based confidence measures. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 101–109
16. 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: 2016 Fourth International Conference on 3D Vision (3DV), IEEE (2016) 509–518
17. Fu, Z., Ardabilian, M., Stern, G.: Stereo matching confidence learning based on multi-modal convolution neural networks. In: International Workshop on Representations, Analysis and Recognition of Shape and Motion From Imaging Data, Springer (2017) 69–81
18. Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S.: Beyond local reasoning for stereo confidence estimation with deep learning. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 319–334
19. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32** (2013) 1231–1237
20. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3061–3070
21. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4040–4048
22. Egnal, G., Mintz, M., Wildes, R.P.: A stereo confidence metric using single view imagery with comparison to five alternative approaches. Image and vision computing **22** (2004) 943–957
23. Mordohai, P.: The self-aware matching measure for stereo. In: 2009 IEEE 12th International Conference on Computer Vision, IEEE (2009) 1841–1848
24. Spyropoulos, A., Mordohai, P.: Correctness prediction, accuracy improvement and generalization of stereo matching using supervised learning. International Journal of Computer Vision **118** (2016) 300–318
25. Veld, R.O.H., Jaschke, T., Bätz, M., Palmieri, L., Keinert, J.: A novel confidence measure for disparity maps by pixel-wise cost function analysis. In: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE (2018) 644–648

26. Kim, S., Min, D., Kim, S., Sohn, K.: Feature augmentation for learning confidence measure in stereo matching. IEEE Transactions on Image Processing **26** (2017) 6019–6033
27. Zhang, Y., Chen, Y., Bai, X., Zhou, J., Yu, K., Li, Z., Yang, K.: Adaptive unimodal cost volume filtering for deep stereo matching. arXiv preprint arXiv:1909.03751 (2019)
28. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 185–194
29. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5410–5418
30. Kim, S., Min, D., Kim, S., Sohn, K.: Unified confidence estimation networks for robust stereo matching. IEEE Transactions on Image Processing **28** (2018) 1299–1313
31. Poggi, M., Tosi, F., Mattoccia, S.: Quantitative evaluation of confidence measures in a machine learning world. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 5228–5237