LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation

Sunok Kim[1,2], Seungryong Kim[1,2], Dongbo Min[3], Kwanghoon Sohn[1*]
[1]Yonsei University [2]École polytechnique fédérale de Lausanne (EPFL) [3]Ewha Womans University
{kso428,khsohn}@yonsei.ac.kr seungryong.kim@epfl.ch dbmin@ewha.ac.kr

Abstract

We present a novel method that estimates confidence map of an initial disparity by making full use of tri-modal input, including matching cost, disparity, and color image through deep networks. The proposed network, termed as Locally Adaptive Fusion Networks (LAF-Net), learns locally-varying attention and scale maps to fuse the tri-modal confidence features. The attention inference networks encode the importance of tri-modal confidence features and then concatenate them using the attention maps in an adaptive and dynamic fashion. This enables us to make an optimal fusion of the heterogeneous features, compared to a simple concatenation technique that is commonly used in conventional approaches. In addition, to encode the confidence features with locally-varying receptive fields, the scale inference networks learn the scale map and warp the fused confidence features through convolutional spatial transformer networks. Finally, the confidence map is progressively estimated in the recursive refinement networks to enforce a spatial context and local consistency. Experimental results show that this model outperforms the state-of-the-art methods on various benchmarks.

1. Introduction

Stereo matching for reconstructing geometric configuration of a scene is one of the fundamental and essential problems in computer vision fields [36]. For decades, numerous methods have been proposed for this task by leveraging handcrafted [43, 10] and/or machine learning based [45, 38] techniques. However, because of its challenging elements such as reflective surfaces, textureless regions, repeated pattern regions, occlusions [23, 13, 6], and photometric deformations incurred by illumination and camera specification variations [44, 9], the stereo matching still remains an unsolved problem. To alleviate these inherent challenges,

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7069370). *Corresponding author

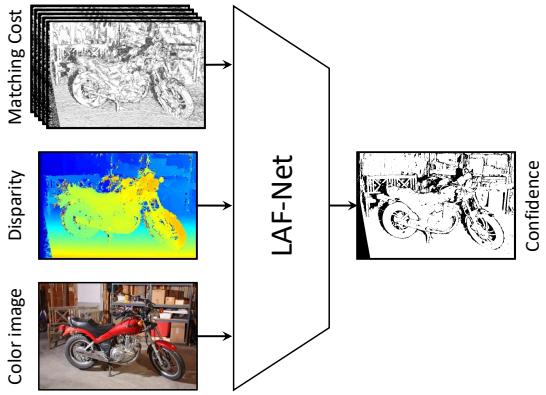


Figure 1. Illustration of LAF-Net: using tri-modal input, consisting of matching cost, disparity, and color image, LAF-Net estimates confidence of disparity.

most methods [39, 27, 29, 20, 18, 21] have adopted the confidence estimation step that detects unreliable disparities and refines them for improving the quality of stereo matching results.

Formally, the confidence estimation pipeline involves first extracting the confidence features and then training the confidence classifiers using ground-truth confidences [39, 27, 30]. Conventionally, there exist several handcrafted confidence measures using different input modalities, such as matching cost, disparity, and color image [12, 28]. Since any single confidence measures cannot handle all failure cases in stereo matching, various combination of hand-designed confidence measures extracted from the tri-modal input [8, 39, 27, 29, 20] has been used to learn shallow classifiers, such as random decision forest [2, 22]. Despite performance improvement by the joint usage of the tri-modal input, they still show a limited performance due to their low discriminative power.

Recent approaches have attempted to estimate the confidence by leveraging deep convolutional neural networks (CNNs) thanks to their high robustness [30, 37, 18, 21], demonstrating the substantial accuracy gain over the handcrafted approaches. However, unlike handcrafted approaches [8, 39, 27] that make full use of the tri-modal input, CNN-based approaches have been formulated by par-

tially using single- or bi-modal input, e.g., matching cost only [38], disparity only [30, 37], matching cost and disparity [18, 21], or disparity and color [7, 40]. Moreover, a simple concatenation technique [16] is commonly used to fuse multi-modal confidence features, disregarding that the fusion weights may vary for each pixel depending on the characteristic of confidence features.

Meanwhile, the receptive fields for confidence features can vary for each pixel. This assumption has been used in conventional handcrafted methods [39, 27, 29, 20] in a way of extracting multi-scale confidence features. For instance, it was reported in [27] that the median disparity deviation value in different scales is the most important confidence features for both outdoor [24] and indoor database [34]. A similar idea has also been adopted in some confidence estimation approaches based on deep CNNs. In [21], the multi-scale disparity feature extraction networks have been proposed to learn the confidence features from disparity in different scales. Also, the dilated convolution that extracts local contextualized information with different dilation factors was proposed by Fu et al. [7]. Tosi et al. [40] proposed local-global confidence networks to effectively combine both local and global context from the input images. However, there is still no mechanism that explicitly considers locally-varying scale fields.

On the other hand, in order to consider the spatial context and local consistency, the output confidence map was refined using joint filtering [20] or using deep CNNs [31], generating more reliable confidence map.

In this paper, we propose novel confidence estimation networks, called Locally Adaptive Fusion Networks (LAFNet), that utilize tri-modal input consisting of matching cost, disparity, and color image as illustrated in Fig. 1. The networks consist of confidence feature extraction networks, attention inference networks, scale inference networks, and recursive confidence refinement networks. In the attention inference networks, we fuse the tri-modal input adaptively with locally-varying attention maps to benefit from the joint usage of the tri-modal confidence features. In the scale inference networks, locally adaptive scale parameters are learned for all pixels, which enables the networks to extract the confidence features within locally optimal receptive fields. In addition, the output confidence is further refined through the recursive confidence refinement networks. The proposed method is extensively evaluated through an ablation study and comparison with conventional handcrafted and CNNs-based methods on various benchmarks, including Middlebury 2006 [34], Middlebury 2014 [33], and KITTI 2015 [24].

2. Related Works

Handcrafted approaches. In last decades, there have been extensive literatures in confidence estimation, mainly based on handcrafted confidence measures [6, 5, 25]. In a comprehensive study of confidence measures has been presented by Hu and Mordohai [12]. Various single confidence measures have been analyzed and categorized according to different input by Park et al. [28]. From matching cost, the peak ratio of the matching costs [11] and naive peak ratio [12] have been widely used to remove unreliable pixels. The maximum margin [12] and winner margin [35] were computed with the difference of matching costs. From disparity, a left-right consistency [5] has been most widely used for finding the correctness of matched pixels. The variances of the disparity (VAR) [8] and the median disparity deviation (MDD) [8] in a local window were also measured to estimate unreliable pixels. Several confidence measures extracted from image have been introduced in [28]. The variance of intensities might be used, especially separating the homogeneous regions from the well-textured regions as well as the magnitude of the image gradients. A distance-to-edge measure incorporated the texturedness of a pixel.

Since there is no single confidence feature that yields stably optimal performance, various approaches to benefit from the feature combination among a different set of single confidence measures have been proposed [8, 39] which trained a shallow classifier such as random decision forest [1, 22]. However, the performance of the aforementioned methods is still limited since the selected confidence features are not optimal. To select the set of (sub-)optimal confidence features among multiple confidence features, Park and Yoon [27] utilized the permutation importance measures to select important set of confidence feautres. In [27], they found the MDD in different scales are important to measure unrelibale pixels. Similarly, Poggi and Mattoccia [29] employed the set of confidence features from only disparity map that can be computed in O(1) complexity without losing the confidence estimation performance. While the aforementioned methods detect unconfident pixels in a pixel-level, Kim et al. [20] leveraged a spatial context to estimate confidence in a superpixel-level. In [20], the resulting confidence map was further refined through hierarchical confidence map aggregation. However, all of these methods used handcrafted confidence features, and they may not be optimal to detect unrelibale pixels on challenging scenes.

Deep CNN-based approaches. Recent approaches have tried to measure the confidence through deep CNNs [30, 37, 31, 18, 21]. A quantitative evaluation of confidence measures that use machine learning approaches has been performed in [32]. Formally, these CNN-based methods first extract the confidence features from single- or bi-modal input and then predict the confidence by jointly learning the feature extractor and classifier. Various methods have been proposed that use the single- or bi-modal input, i.e., a left disparity [30], both left and right disparity [37], a matching

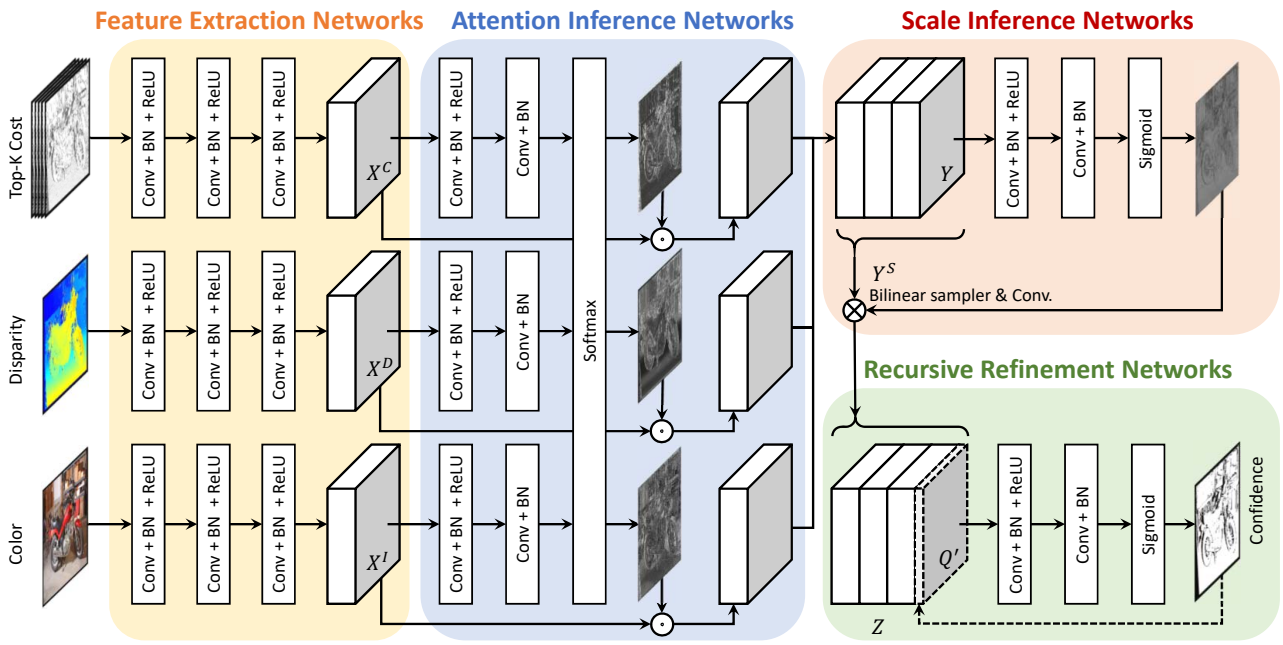


Figure 2. The network configuration of LAF-Net which consists of four sub-networks, including feature extraction networks, attention inference networks, scale inference network, and recursive refinement networks. Given matching cost, disparity, and color image as input, our networks output confidence of the disparity. The detail of scale inference network is illustrated in Fig. 4.

cost [38], matching cost and disparity [18, 21], and disparity and color [7, 40]. In order to extract confidence features from matching cost and disparity, Kim et al. [21] proposed the top-K pooling layer to normalize matching cost and improved the discriminative power to classify the unrelibale pixels. Although these methods improved the confidence estimation performance, they did not make full use of the tri-modal input.

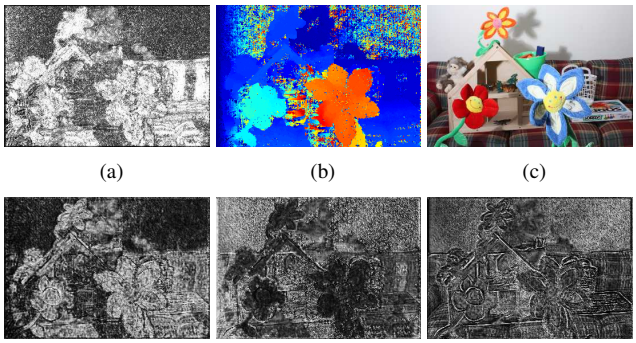
In [21], the multi-scale disparity feature extractor was proposed, while dilation convolution was applied in [7] to gain local contexualized information effectively. In [40], they proposed global confidence measure using encoder-decoder networks by looking at the whole image and disparity content. By using the output of global confidence, they proposed local-global approach by fusing the local confidence, the global confidence, and disparity. All of these methods considered only fixed and pre-defined scale ranges and did not estimate a scale that varying for each pixel. On the other hand, the confidence refinement networks [31] were also developed, which can improve the accuracy of the estimated confidence map by leveraging a local consistency within the confidence map.

3. Proposed Method

3.1. Problem Statement and Motivation

Let us define a pair of stereo images as I^l and I^r, respectively. The objective of stereo matching is to estimate a disparity D_i between the stereo image pairs that is defined for each pixel $i = [i_x, i_y]^T$. The matching costs $C_{i,d}$ between I_i^l and $I_{i'}^r$, where $i' = i - [d, 0]^T$, among disparity candidates $d = \{1, ..., d_{\max}\}$ are first measured, and then aggregated and optimized for computing the disparity D_i. Most existing methods for stereo matching [10, 17, 45] cannot provide fully reliable results due to its challenging elements, thus several approaches [39, 27, 37, 18, 20] have presented an additional module to predict a confidence Q_i of the disparity D_i. By leveraging the confidence Q_i, they refine the initial disparity D_i through subsequent disparity refinement pipeline.

To realize this, we design a novel network architecture that estimates the confidence by fully exploiting matching cost C, disparity map D, and color image I. The overall networks consist of four sub-networks, including confidence feature extraction networks, attention inference networks, scale inference networks, and recursive confidence refinement networks, as illustrated in Fig. 2. In feature extraction networks, confidence features are first extracted from tri-modal input. The intermediate features from this network are then fed to learn locally-varying attention maps in attention inference networks. The attention maps are used to adaptively concatenate the tri-modal confidence features, unlike existing approaches [18, 21, 7, 40] that use a simple concatenation technique. Then, locally-varying scale fields are learned for extracting confidence features within geometrically-aligned receptive fields through scale inference networks, different from conventional approaches [30, 37, 21] with a fixed-size convolution.



(a) (b) (c) (d) (e) (f)

Figure 3. The visulization of the attention maps: (a) top-1 matching cost, (b) initial disparity, (c) left color image, (d)-(f) the attention maps for matching cost, disparity, and color, respectively.

Finally, the confidence is progressively refined in recursive confidence refinement networks to enforce a spatial context and local consistency inspired by [20, 31].

3.2. Confidence Feature Extraction Networks

The confidence feature extraction networks are designed to extract the tri-modal confidence features denoted as X^C, X^D, and X^I from matching cost C, disparity D, and left color image[1] I^l by feed-forward processes such that $X^C = \mathcal{F}(C; W^C)$, $X^D = \mathcal{F}(D; W^D)$, and $X^I = \mathcal{F}(I^l; W^I)$ with network parameters W^C, W^D, and W^I, respectively. The network parameters for each network are seperately learned, not shared, to encode the heterogeneous charateristics of the tri-modal input. The size and absolute value of raw matching cost C_{raw} vary depending on the search range of stereo image pairs and stereo matching methods. Additionally, its distribution is often non-discriminative as mentioned in [37, 7]. To alleviate these limitations, the input matching cost C_{raw} is converted into a top-K matching probabiliry[2] C as in [18, 21], which enables the search range-invariant convolutions.

The confidence feature extraction networks consist of 3 convolutional layers (Conv) with 3×3 kernels producing 64 feature channel, followed by batch normalization (BN) and rectified linear units (ReLU).

3.3. Attention Inference Networks

Due to their heterogeneous attributes, a direct concatenation of these tri-modal input does not provide an optimal performance [7]. Alternatively, some methods [18, 7, 40, 21] first extract the bi-modal confidence features and then concatenate them. However, such a simple approach that fixes the fusion weights at inference often fails to perform an optimal feature fusion.

To alleviate this limitation, inspired by [15], we build the attention inference networks for inferring an optimal fusion weight among the tri-modal features, i.e., X^C, X^D, and X^I. The locally-varying attention for each modality at pixel i is defined as A_i^C, A_i^D, and A_i^I for matching cost, disparity, and color image, respectively. These attentions are learned such that $A_i^C = \mathcal{F}(X_i^C; W_C^A)$, $A_i^D = \mathcal{F}(X_i^D; W_D^A)$, and $A_i^I = \mathcal{F}(X_i^I; W_I^A)$ with the network parameters W_C^A, W_D^A, and W_I^A, and these attentions then undergoes a softmax function to make the sum of attentions for each pixel to be 1, i.e., $\sum_{*\in C,D,I}(A_i^*) = 1$. Note that the attention inference network parameters for each modality (i.e., W_C^A, W_D^A, and W_I^A) are not shared but independently learned depending on their attributes.

The learned attentions are then applied to the confidence feature as

$$Y_i = \Pi\left(X_i^C \odot A_i^C, X_i^D \odot A_i^D, X_i^I \odot A_i^I\right), \qquad (1)$$

where $\Pi(\cdot)$ is a concatenation operator and $\odot$ is an element-wise multiplication operator. Note that unlike methods [7, 21, 40] that use the fixed fusion weights, the attentions A^C, A^D, and A^I, are estimated conditioned on input and varies locally, thus enabling the data-adaptive fusion more effectively. The visualization of attention maps for different input modalities is exemplified in Fig. 3. The attention of top-K matching cost is high for pixels having high matching probability. On the other hand, the attention of disparity has high value in noisy region, indicating informative features can be extracted from the differnet disparity assignments, as considered similar to VAR or MDD [8] in handcafted features. In color image, the attentions near image boundary are high and this indicates a image texture can give a useful cue to estimate confidence. By adaptively weighting the confidence features with these attention maps, we can obtain more discriminative confidence features.

The attention learning networks consist of 2 Conv with 3×3 kernels. The first Conv produces 64 channel feature, followed by BN and ReLU, and the second Conv produces 1 channel feature followed by only BN.

3.4. Scale Inference Networks

The optimal receptive fields for confidence features can vary at each pixel. In order to encode confidence features of different scales, some approaches [27, 7, 21, 40] have been proposed, but they consider only fixed and pre-defined scale ranges and do not estimate scales that vary for each pixel.

To determine the optimal receptive fields for confidence features at each pixel, we present the scale inference networks that learn locally-varying scale fields. It first infers the scale fields through subsequent convolutions such that $S_i = \mathcal{F}(Y_i; W^S)$ with network parameters W^S. With these scale fields S_i, the intermediate features are warped through

[1] We use a left color image only to estimate the confidence of left disparity and a right image can be used when estimating the confidence of right disparity.

[2] We denote this as the matching cost for the sake of clarity.

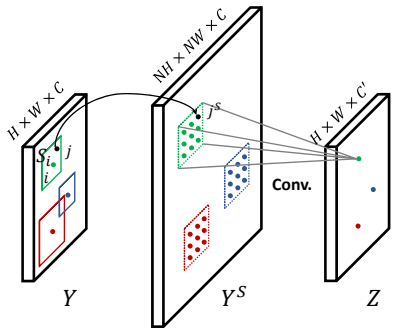


Figure 4. Illustration of a bilinear sampler in the scale inferent networks: for each pixel i in the feature Y can be warped as enlarged size feature Y^S. The neighbors j^S is convolved as Z with stride.

an image sampling on a parameterized grid, similar to spatial transformer networks (STNs) [14].

However, a spatially-varying parameterized sampling grid cannot be directly realized with the original STNs [14] that is designed for a global geometric field. To deal with locally-varying scale fields, we first build a locally-varying sampling grid for $N \times N$ neighbors $j \in \mathcal{N}_i$ indenpendently, and then warp the convolutional activation for each sampling grid as used in [4, 19]. Concretely, the locally-varying sampling grid $j^S = [j^S_{\mathbf{x}}, j^S_{\mathbf{y}}]^T$ is defined such that

$$\begin{bmatrix} j^S_{\mathbf{x}} \\ j^S_{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} S_i & 0 \\ 0 & S_i \end{bmatrix} \begin{bmatrix} j_{\mathbf{x}} - i_{\mathbf{x}} \\ j_{\mathbf{y}} - i_{\mathbf{y}} \end{bmatrix} + \begin{bmatrix} i_{\mathbf{x}} \\ i_{\mathbf{y}} \end{bmatrix}, \quad (2)$$

for all pixels i and their neighbors j within receptive fields on the regular grid. For each grid sample $j^S = [j^S_{\mathbf{x}}, j^S_{\mathbf{y}}]^T$, receptive fields for convolutional layers are warped through the bilinear sampler [14] independently such that

$$Y^S_{i,j} = \sum_i Y_i \max(0, 1 - |j^S_{\mathbf{x}} - i_{\mathbf{x}}|) \max(0, 1 - |j^S_{\mathbf{y}} - i_{\mathbf{y}}|), \quad (3)$$

where $Y^S_{i,j}$ is the warped convolutional activation of $Y_{i,j}$. Since this scale-varying convolutional features are defined for all i and j independently, the spatial size of Y^S is enlarged as $|\mathcal{N}|$ times of the size of Y without overlap as illustrated in Fig. 4. Then, Y^S passes through a subsequent convolution with the stride N to convolve the warped features independently and generate the scale-adaptive confidence features Z. We chose N as 3 since the kernel size of following convolutional layer is 3×3.

The scale learning networks consist of 2 Conv with 3×3 kernels. The first Conv produces 64 channel feature, followed by BN and ReLU, and the second Conv produces 1 channel feature followed by only BN. The output passes through the sigmoid layer to generate the scale parameter for each pixel.

3.5. Recursive Confidence Refinement Networks

So far we introduce our networks that fuse tri-modal confidence features through the attention and scale inference networks. From the confidence feature Z_i, we finally formulate the confidence prediction networks to estimate the confidence Q_i such that $Q_i = \mathcal{F}(Z_i; W^P)$ with the prediction parameters W^P. The iterative refinement procedure of output confidence can improve the confidence estimation accuracy as studied in the handcrafted approach using joint filtering [20] and CNNs-based approach [31]. Inspired by this, we propose the recursive confidence refinement networks, where the previously estimated confidence serves as a guidance of the current confidence estimation. To realize this recursive module, we formulate the networks such that $Q^t_i = \mathcal{F}(Z_i, Q^{t-1}_i; W^P)$ where Q^t_i and Q^{t-1}_i are the estimated confidences at t^{th} and $(t-1)^{\text{th}}$ iteration, respectively. The initial confidence Q^0_i is defined as zeros. As evolving the iterations, the confidence accuracy is improved gradually and the final confidence map is obtained as $Q' = Q^{t_{\max}}$. The effectiveness of the recursive confidence refinement networks is shown in Fig. 5. Here, we set 0.9 to threshold. With the recursive module, the ability to predict mismatched pixels on initial disparity is improved.

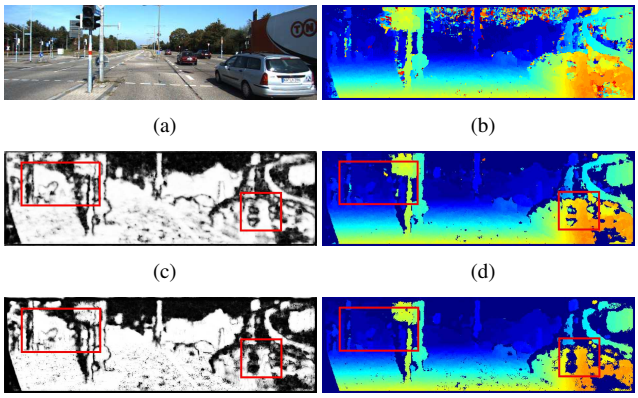


Figure 5. The effectiveness of the proposed recursive confidence refinement networks: (a) left color image, (b) initial disparity, (c) estimated confidence map without recursive module, (d) thresholded disparity with (c), (e) estimated confidence map with recursive module, (f) thresholded disparity with (e). The mismatched pixels in the red boxes are reliably detected with the proposed recursive confidence refinement networks.

The recursive confidence refinement networks consist of 2 Conv and final sigmoid layer similar to the scale learning networks. For the number of iteration, we set $t_{\max}$ to 3.

The proposed method employs the cross-entropy loss function [38, 21] with respect to the ground-truth confidence Q^* and the estimated confidence Q'_i.

4. Experimental Results

4.1. Experimental Settings

The proposed method was implemented in MATLAB with `VLFeat MatConvNet toolbox` [42] and simu-

Match. cost	✓		✓		✓
Disparity		✓		✓	✓
Color			✓	✓	✓
MID 2006	0.0431	0.0392	0.0381	0.0375	**0.0364**
MID 2014	0.0762	0.0703	0.0687	0.0685	**0.0683**
KITTI 2015	0.0347	0.0245	0.0237	0.0231	**0.0225**

Table 1. Ablation study for the various combination of input modalities in LAF-Net on MID 2006 [34], MID 2014 [33], and KITTI 2015 [24] dataset, when the raw matching cost is obtained using MC-CNN [45].

Attention	✓			✓	✓
Scale		✓		✓	✓
Recursive			✓		✓
MID 2006	0.0374	0.0375	0.0372	0.0371	**0.0364**
MID 2014	0.0686	0.0688	0.0685	0.0685	**0.0683**
KITTI 2015	0.0235	0.0236	0.0231	0.0229	**0.0225**

Table 2. Ablation study for the effectivness of each sub-networks in LAF-Net on MID 2006 [34], MID 2014 [33], and KITTI 2015 [24] dataset, when the raw matching cost is obtained using MC-CNN [45]. The average AUC values for simple concatenation without fusion methods are 0.0386, 0.0689, and 0.0238 for MID 2006, MID 2014, and KITTI 2015, respectively.

lated on a PC with TitanX GPU. We make use of the stochastic gradient descent with momentum, and set the learning rate to 1×10^{-6} and the batch size to 16. To compute a raw matching cost, we used a census transform with a 5×5 local window and MC-CNN [45], respectively. For the census transform, we applied SGM [10] on estimated cost volumes by setting $P_1 = 0.008$ and $P_2 = 0.126$ as in [27]. For computing the MC-CNN, 'KITTI 2012 fast network' was used, provided at the author's website [46]. We set σ as 100 and 0.05 for census-SGM and MC-CNN, respectively, as in [21]. We trained our networks using MPI Sintel dataset [3] and KITTI 2012 dataset [24], and evaluated each model on Middlebury 2006 (MID 2006) [34], Middlebury 2014 (MID 2014) [33], and KITTI 2015 dataset [24]. In addition, we used the half-sized KITTI database due to memory constraints, so we measured the error rates and AUC values in the half-sized resolution. For Middlebury, we used the third-sized images provided by [34]. The ground-truth confidence maps are obtained by thresholding an absolute difference between estimated disparity and ground-truth disparity to 1. In inference, the LAF-Net takes about 0.912s, 2.413s, and 0.783s for MID 2006 (368×424), MID 2014 (496×792), and KITTI 2015 (608×184), respectively, while [40] takes 0.750s, 1.628s, and 0.552s in the same settings. Due to the bilinear sampler and recursive procedure, the LAF-Net takes longer than [40]. In contrast, the number of parameters in LAF-Net and [40] is 1,337K and 9,289K, proving that LAF-Net is lighter while achieving a better accuracy.

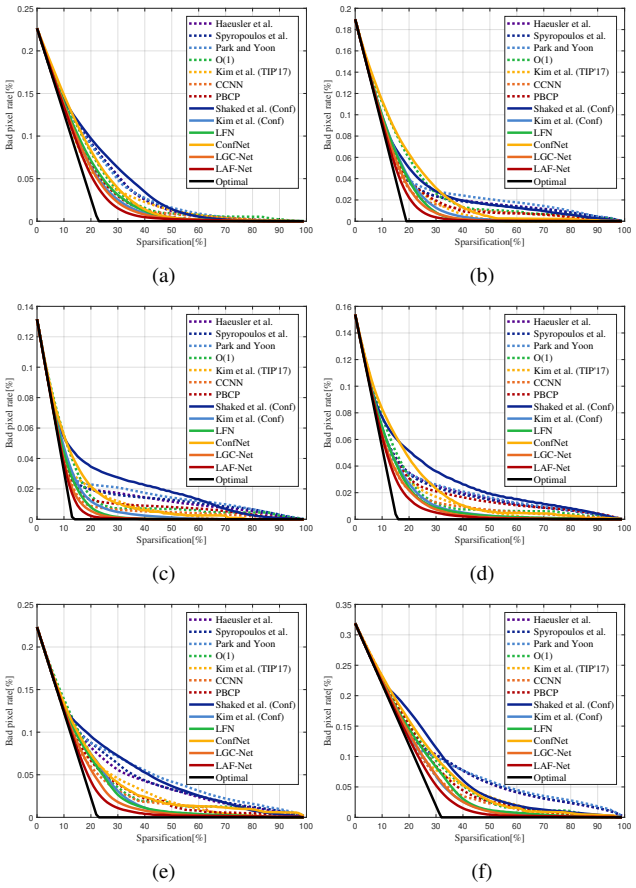


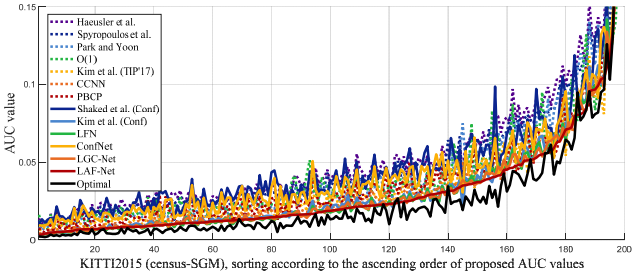
Figure 6. The sparsification curves of selected images for MID 2006 [34], MID 2014 [33], and KITTI 2015 dataset [24] using (a), (c), (e) census-SGM and (b), (d), (f) MC-CNN. The sparsification curve for the ground-truth confidence map is described as 'optimal'.

In the following, we evaluated the proposed method in comparison to conventional handcrafted approaches, such as Haeusler et al. [8], Spyropoulos et al. [39], Park and Yoon [27], Poggi and Mattoccia [29], Kim et al. [20]. Several CNNs-based approaches using single- or bi-modal input are also compared, where using disparity only, such as Poggi and Mattoccia (CCNN) [30], Seki and Pollefey (PBCP) [37], matching cost only, such as Shaked et al. [38], both disparity and matching cost, such as Kim et al. [21], and both color and disparity, such as Fu et al. (LFN) [7] and the global measures of Tosi et al. (ConfNet) [40] and local and global measures (LGC-Net) [40]. We obtained the results of [27], [20], and [21] by using the author-provided code, while the results of [8], [39], [37], [38], and [7] were obtained by our own implementation. We re-implemented methods of [29], [30], and [40] based on the author-provided code.

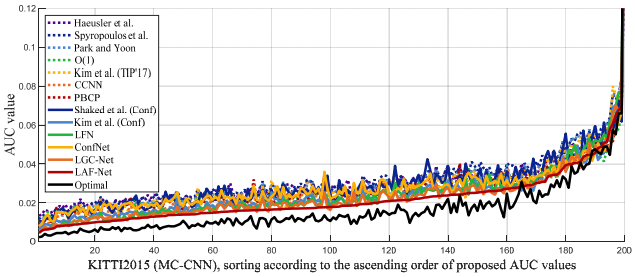
To evaluate the performance of confidence estimation quantitatively, we used the sparsification curve and its area under curve (AUC) as used in [8, 39, 27, 37, 21]. The spar-

Datasets	MID 2006 [34]		MID 2014 [33]		KITTI 2015 [24]	
	Census-SGM	MC-CNN	Census-SGM	MC-CNN	Census-SGM	MC-CNN
Haeusler et al. [8]	0.0454	0.0417	0.0841	0.0750	0.0585	0.0308
Spyropoulos et al. [39]	0.0447	0.0420	0.0839	0.0752	0.0536	0.0323
Park and Yoon [27]	0.0438	0.0426	0.0802	0.0734	0.0527	0.0303
Poggi et al. [29]	0.0439	0.0413	0.0791	0.0707	0.0461	0.0263
Kim et al. [20]	0.0430	0.0409	0.0772	0.0701	0.0430	0.0294
CCNN [30]	0.0454	0.0402	0.0769	0.0716	0.0419	0.0258
PBCP [37]	0.0462	0.0413	0.0791	0.0718	0.0439	0.0272
Shaked et al. (Conf) [38]	0.0464	0.0495	0.0806	0.0736	0.0531	0.0292
Kim et al. (Conf) [21]	0.0419	0.0394	0.0749	0.0694	0.0407	0.0250
LFN [7]	0.0416	0.0393	0.0752	0.0692	0.0405	0.0253
ConfNet [40]	0.0451	0.0428	0.0783	0.0721	0.0486	0.0277
LGC-Net [40]	0.0413	0.0389	0.0735	0.0685	0.0392	0.0236
LAF-Net	**0.0405**	**0.0364**	**0.0718**	**0.0683**	**0.0385**	**0.0225**
Optimal	0.0340	0.0323	0.0569	0.0527	0.0348	0.0170

Table 3. The average AUC values for MID 2006 [34], MID 2014 [33], and KITTI 2015 [24] dataset. The AUC value of ground truth confidence is measured as 'Optimal'. The result with the lowest AUC value in each experiment is highlighted.



(a)



(b)

Figure 7. Comparisons of AUC values for (a) census-based SGM and (b) MC-CNN for the KITTI 2015 dataset [24]. We sort the AUC values in the ascending order according to the AUC values.

sification curve draws a bad pixel rate while successively removing pixels in descending order of confidence values in the disparity map, thus it enables us to observe the tendency of estimation errors. For the higher accuracy of the confidence measure, AUC value is lower and the optimal AUC is measured using ground-truth confidence.

4.2. Ablation Study

We analyzed our confidence estimation networks with the ablation evaluations, with respect to various combination of different modalities and the effectiveness of the proposed sub-networks.

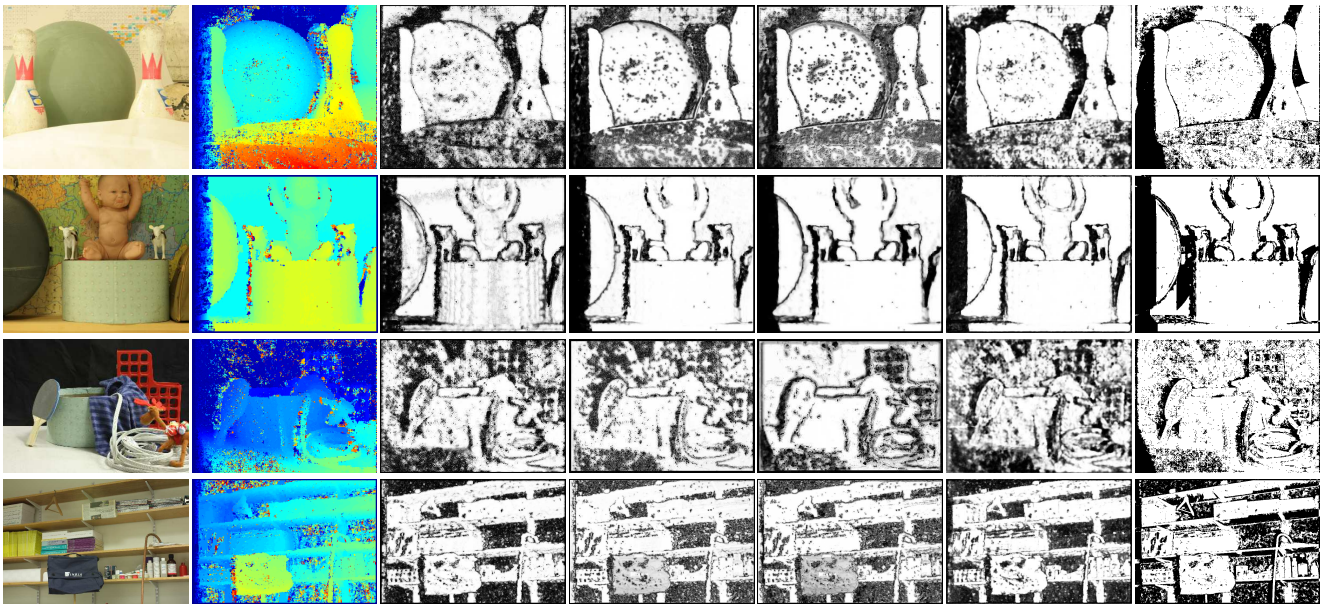
The effects on tri-modal input. In Table 1, ablation experiments to validate the effects of multi-modal input show the necessity of using the tri-modal input. Note that the attention inference module is not used for input of single modality. Although the bi-modal input improved the ability to predict reliable pixels, the full usage of tri-modal input shows the best performance.

The effects on various fusion methods. In Table 2, ablation experiments to validate the effects of the proposed fusion methods. Compared to the simple concatenation technique, the confidence estimator is improved with the attention and scale obtained from the attention and scale inference networks. Also, the recursive confidence refinement networks show the additional improvement.

4.3. Confidence Estimation Analysis

In order to measure the performance of the confidence estimator in comparison to other methods, we compared the average AUC values of our method with conventional learning-based approaches using handcrafted confidence measures [8, 39, 27, 29, 20] and CNNs-based methods [37, 30, 7, 40]. For fair comparison, we also evaluated the confidence estimation performance only for [38, 21], i.e., Shaked et al. (Conf) [38] and Kim et al. (Conf) [21].

Sparsification curves for MID 2006 [34], MID 2014 [33], and KITTI 2015 [24] with census-based SGM and MC-CNN are shown in Fig. 6. Fig. 7 describes the AUC values, which are sorted in ascending order, for the KITTI 2015 [24] with census-based SGM and MC-CNN, respectively. The results have shown that the proposed confidence estimator exhibits a better performance than both



(a) (b) (c) (d) (e) (f) (g)

Figure 8. The confidence maps on MID 2006 dataset [34] (first two rows) and MID 2014 dataset [33] (last two rows) using census-SGM and MC-CNN. (a) color images, (b) initial disparity map, (c)-(f) are estimated confidence maps by (c) Kim et al. [21], (d) LFN [7], (e) LGC-Net [40], (f) LAF-Net, and (g) ground-truth confidence map.

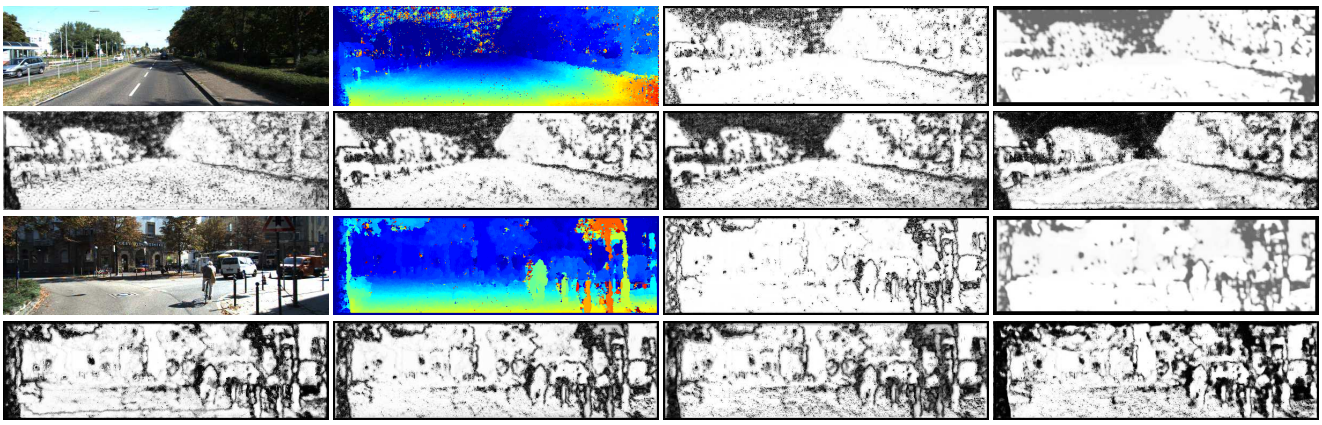


Figure 9. The confidence maps on KITTI 2015 dataset [24] using census-SGM (first two rows), and MC-CNN (last two raws). (From top to bottom, left to right) color images, initial disparity map, estimated confidence maps by CCNN [30], PBCP [37], Kim et al. [21], LFN [7], LGC-Net [40], and LAF-Net.

conventional handcrafted approaches and CNN-based approaches. The average AUC with census-based SGM and MC-CNN for MID 2006, MID 2014, and KITTI 2015 datasets were summarized in Table 3. The handcrafted approaches showed inferior performance than the proposed method due to low discriminative power. CNNs-based methods [30, 37, 38, 7] have improved confidence estimation performance compared to existing handcrafted methods such as [8, 39, 27, 29, 20], but they are still inferior to our method as they rely on single- [30, 38] or bi-modal [37, 21, 7, 40] input rather than tri-modal input. The estimated confidence maps are shown in Fig. 8 and Fig. 9.

5. Conclusion

We presented LAF-Net that estimates confidence with tri-modal input, including matching cost, disparity, and color image through deep networks. The key idea of the proposed method is to design locally adaptive attention and scale inference networks to generate optimal fusion weights. In addition, the confidence estimation performance is further improved with recursive confidence refinement networks. A direction for further study is to examine how confidence estimation networks could be learned in an unsupervised manner as proposed in [26, 41].

References

[1] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725, 2000.

[2] L. Breiman. Random forests. *Mach. Learn.*, 63(4):5–32, 2001.

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. *in Proc. Eur. Conf. Comput. Vis.*, pages 611–625, Oct. 2012.

[4] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. *in Proc. Advances in Neural Inf. Process. Syst.*, pages 2414–2422, Dec. 2016.

[5] G. Egnal, M. Mintz, and R. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image. Vis. Comput.*, 22(12):943–957, 2004.

[6] G. Egnal and R. P. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(8):1127–1133, 2002.

[7] Z. Fu and M. A. Fard. Learning confidence measures by multi-modal convolutional neural networks. *in Proc. IEEE Winter Conf. Applicat. Comput. Vis.*, pages 1321–1330, 2018.

[8] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measrues in stereo vision. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 305–312, Jun. 2013.

[9] Y. Heo, K. Lee, and S. Lee. Robust stereo matching using adaptive normalized cross correlation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4):807–822, 2011.

[10] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008.

[11] H. Hirschmuller, P. Innocent, and J. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *Int. J. Comput. Vis.*, 47(1–3):229–246, 2002.

[12] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2121–2133, 2012.

[13] M. Humenberger, C. Zinner, M. Weber, W. Kubinger, and M. Vincze. A fast stereo matching algorithm suitable for embedded real-time systems. *Comput. Vis. Image. Understand.*, 114(11):1180–1202, 2010.

[14] M. Jaderberg, K. Simonyan, and A. Zisserman. Spatial transformer networks. *in Proc. Advances in Neural Inf. Process. Syst.*, pages 2017–2025, Dec. 2015.

[15] X. Jia, B. D. Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. *in Proc. Advances in Neural Inf. Process. Syst.*, pages 667–675, Dec. 2016.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1725–1732, Jun. 2014.

[17] S. Kim, B. Ham, B. Kim, and K. Sohn. Mahalanobis distance cross-correlation for illumination invariant stereo matching. *IEEE Trans. Circ. Syst. Vid. Techn.*, 24(11):1844–1859, 2014.

[18] S. Kim, D. Min, B. Ham, S. Kim, and K. Sohn. Deep stereo confidence prediction for depth estimation. *in Proc. IEEE Conf. Image. Process.*, Sep. 2017.

[19] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[20] S. Kim, D. Min, S. Kim, and K. Sohn. Feature augmentation for learning confidence measure in stereo matching. *IEEE Trans. Image Process.*, 26(12):6019–6033, 2017.

[21] S. Kim, D. Min, S. Kim, and K. Sohn. Unified confidence estimation networks for robust stereo matching. *IEEE Trans. Image Process.*, 28(3):1299–1313, 2019.

[22] A. Liaw and M. Wiener. Classification and regression by random forest. *R news*, 2(3):18–22, 2002.

[23] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. *in Proc. IEEE Int. Conf. Comput. Vis. Work.*, pages 467–474, Nov. 2011.

[24] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3061–3070, Jun. 2015.

[25] P. Mordohai. The self-aware matching measure for stereo. *in Proc. IEEE Int. Conf. Comput. Vis.*, pages 1841–1848, Sep. 2009.

[26] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Using self-contradiction to learn confidence measures in stereo vision. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4067–4076, Jun. 2016.

[27] M. Park and K. Yoon. Leveraging stereo matching with learning-based confidence measures. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 101–109, Jun. 2015.

[28] M. Park and K. Yoon. Learning and selecting confidence measures for robust stereo matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[29] 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 Proc. IEEE Int. Conf. 3D Vis.*, pages 509–518, Oct. 2016.

[30] M. Poggi and S. Mattoccia. Learning from scratch a confidence measure. *in Proc. Brit. Mach. Vis. Conf.*, 10, Sep. 2016.

[31] M. Poggi and S. Mattoccia. Learning to predict stereo reliability enforcing local consistency of confidence maps. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017.

[32] M. Poggi, F. Tosi, and S. Mattoccia. Quantitative evaluation of confidence measures in a machine learning world. *in Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017.

[33] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. *in Proc. German Conf. Pattern Recognit.*, pages 31–42, Sep. 2014.

[34] D. Scharstein and C. Pal. Learning conditional random fields for stereo. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, Jun. 2007.

[35] D. Scharstein and R. Szeliski. Stereo matching with non-linear diffusion. *Int. J. Comput. Vis.*, 28(2):155–174, 1998.

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

[37] A. Seki and M. Pollefeys. Patch based confidence prediction for dense disparity map. *in Proc. Brit. Mach. Vis. Conf.*, 10, Sep. 2016.

[38] A. Shaked and L. Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017.

[39] A. Spyropoulos, N. Komodakis, and P. Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1621–1628, Jun. 2014.

[40] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. *in Proc. Eur. Conf. Comput. Vis.*, Sep. 2018.

[41] F. Tosi, M. Poggi, A. Tonioni, L. D. Stefano, and S. Mattoccia. Learning confidence measures in the wild. *in Proc. Brit. Mach. Vis. Conf.*, page 2, Sep. 2017.

[42] A. Vedaldi and K. Lnc. Matconvnet: Convolutional neural networks for matlab. *in Proc. ACM Int. Conf. Multimedia*, pages 689–692, Oct. 2015.

[43] K. J. Yoon and I. S. Kweon. Adaptive support-weight approach for correspondence search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):650–656, 2006.

[44] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. *in Proc. Eur. Conf. Comput. Vis.*, pages 151–158, May 1994.

[45] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1592–1599, Jun. 2015.

[46] 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.