Unified Confidence Estimation Networks for Robust Stereo Matching

Sunok Kim, *Student Member, IEEE*, Dongbo Min, *Senior Member, IEEE*, Seungryong Kim, *Member, IEEE*, and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*— We present a deep architecture that estimates a stereo confidence, which is essential for improving the accuracy of stereo matching algorithms. In contrast to existing methods based on deep convolutional neural networks (CNNs) that rely on only one of the matching cost volume or estimated disparity map, our network estimates the stereo confidence by using the two heterogeneous inputs simultaneously. Specifically, the matching probability volume is first computed from the matching cost volume with residual networks and a pooling module in a manner that yields greater robustness. The confidence is then estimated through a unified deep network that combines confidence features extracted both from the matching probability volume and its corresponding disparity. In addition, our method extracts the confidence features of the disparity map by applying multiple convolutional filters with varying sizes to an input disparity map. To learn our networks in a semi-supervised manner, we propose a novel loss function that use confident points to compute the image reconstruction loss. To validate the effectiveness of our method in a disparity post-processing step, we employ three post-processing approaches; cost modulation, ground control points-based propagation, and aggregated ground control points-based propagation. Experimental results demonstrate that our method outperforms state-of-the-art confidence estimation methods on various benchmarks.**

***Index Terms*— Stereo confidence, confidence learning, matching probability volume, confidence estimation network.**

I. Introduction

FOR decades, the stereo matching has been one of the fundamental and essential topics in the fields of computer vision. It aims to estimate accurate corresponding points for each pixel between a pair of two images taken under different viewpoints of the same scene. Though numerous methods have been proposed for this task, it still remains an unsolved problem due to several factors including textureless or repeated pattern regions, and occlusions [1]–[3]. Besides, photometric deformations incurred by illumination and/or camera variations pose additional challenges [4]–[6].

In general, most of the stereo matching methods consist of the following pipelines [7]: 1) matching cost computation, 2) matching cost aggregation, 3) disparity regularization, and 4) post-processing. Concretely, the similarity between patches from left and right images is first measured with various cost measures [8], [9] and the matching cost volume is then constructed with these matching costs across disparity candidates. Various methods have attempted to improve the matching accuracy at each step. To robustly compute the matching cost, several approaches were proposed by using robust cost measures [6], [10], [11] or learning-based approaches [9], [12]. Moreover, the matching costs are aggregated to alleviate matching ambiguities by considering local neighbors [13]–[15]. Powerful regularization techniques [8], [16] can also be applied by incorporating prior constraints into an objective function. These approaches help to yield reliable matching results to some extent, but they do not fully address the inherent problems of stereo matching.

To further improve the matching accuracy, most approaches involve the post-processing step. A set of unreliable pixels is first extracted using confidence measures, and then interpolated using reliable estimates at neighboring pixels [8], [13], [17]–[20]. Conventionally, mismatched pixels were detected using simple confidence measures such as a left-right consistency or peak ratio [21], [22]. Recently, learning-based approaches have been popularly proposed to boost the performance of confidence estimation [20], [23]–[25]. Formally, a set of confidence features extracted from stereo matching results of the training data is used to train the confidence classifier [20], [23], [26]. In a test phase, the trained classifier is used to estimate the confidence of each pixel. These methods have shown distinct strengths in comparison to existing non-learning based approaches [21]. However, they still rely on hand-crafted confidence features for training the confidence classifier, and thus they often fail to detect mismatched pixels under challenging conditions [27].

More recently, convolutional neural networks (CNNs)-based approaches have been proposed to robustly extract confidence features from a disparity map and estimate the confidence [28], [29]. Although they have shown improved performance than the existing methods based on handcrafted features [20], [23], [24], [30], they rely on rather limited information in learning the confidence features. According to [31], there exist

Manuscript received February 9, 2018; revised July 13, 2018 and September 3, 2018; accepted October 21, 2018. Date of publication October 26, 2018; date of current version November 7, 2018. This work was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grant NRF-2017M3C4A7069370. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ce Zhu. *(Corresponding author: Kwanghoon Sohn.)*

S. Kim, S. Kim, and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: kso428@yonsei.ac.kr; srkim89@yonsei.ac.kr; khsohn@yonsei.ac.kr).

D. Min is with the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea (e-mail: dbmin@ewha.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2878325

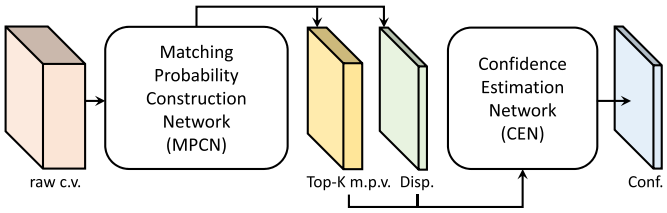


Fig. 1. Illustration of the proposed network, consisting of two sub-networks including matching probability volume construction network and confidence estimation network.

many informative features within the cost volume. It was also shown in many literatures [20], [23], [24], [30] that the hand-crafted features extracted from the cost volume can boost the performance of the confidence estimation. In contrast, most of the existing CNN-based approaches leverage only single or stereo disparity maps for confidence estimation [28], [29], which may limit the performance in challenging conditions. Although several methods [28], [32] have attempted to use the cost volume in the confidence estimation within deep networks, the direct usage of the raw cost volume cannot boost the confidence estimation performance. Unlike disparity information, the raw matching cost volume has scale variation problem that the absolute value of the matching cost volume varies depending on stereo matching approaches used and its distribution is non-discriminative for reliable and unreliable pixels. This prohibits the direct usage of the cost volume in the confidence estimation networks.

In this paper, we propose a novel deep architecture that simultaneously uses a matching cost volume and disparity map as inputs to estimate the confidence as in Fig. 1. Firstly, we propose a matching probability construction network (MPCN) that extracts the matching probabilities from the cost volume to improve a discriminative power and solve the scale variation problem of the cost volume. Secondly, we propose a confidence estimation network (CEN) to estimate the reliable confidence by simultaneously leveraging the matching probabilities and their corresponding disparity maps, where multiple convolutional filters with varying sizes are used to extract multi-scale disparity features. These two sub-networks are learned in a joint and boosting manner, enabling us to estimate reliable disparity and confidence maps in a unified framework.

To learn our networks in a supervised manner, vast amount of stereo image pairs and corresponding dense disparity (or depth) maps are required. However, in outdoor scenes, 3D laser scanners often fail to capture depth information of very distant objects. For instance, no depth information is provided at the upper parts (e.g. sky) of color images in the KITTI benchmark [33], making supervised learning of CNNs infeasible for this task. Alternatively, dense ground truth disparity maps would be generated through synthetic rendering, as in the MPI sintel [34] and DispNet [35], but they cause domain adaptation problem inherently when directly used to estimate disparity maps for real outdoor stereo images. To address the problem of limited training samples for the stereo matching, we propose a semi-supervised learning scheme in which the pixels classified as confident through the estimated confidence map are used to measure the image reconstruction loss [36].

To verify the proposed confidence estimation method, we employ three post-processing methods such as cost modulation [24], ground control points (GCPs)-based propagation [37], and aggregated GCPs-based propagation [38]. Experimental results show that the proposed method outperforms conventional handcrafted feature-based methods and CNN-based methods on various benchmarks, including Middlebury 2006 [39], Middlebury 2014 [40], and KITTI 2015 [33].

This manuscript extends the conference version of this work [41]. It newly adds (1) additional convolutional networks to refine the raw matching cost volume to boost confidence estimation performance; (2) a semi-supervised learning scheme to learn the proposed networks only with the limited sparse ground truth disparity maps; (3) multi-scale disparity feature extraction networks to extract confidence features from different scale of disparities; and (4) an extensive comparative study with the state-of-the-art confidence estimation algorithms using various datasets.

II. Related Works

A. Hand-Crafted Approaches

Numerous approaches have been proposed for stereo confidence estimation based on various handcrafted confidence measures [3], [17], [18]. A comprehensive review provided in [31] concluded that using single confidence feature would not yield good performance in confidence estimation. In order to overcome such limitations and improve the prediction accuracy, there have been various attempts to combine confidence features and train a simple shallow classifier, e.g. random decision forest [20], [23]–[25]. They define the confidence feature, which consists of various stereo confidence cues from the estimated disparity map and (optionally) cost volume. Haeusler *et al*. [23] combined confidence features consisting of left-right consistency, image gradient, and disparity map variance for training the classifier. Similar approach was also proposed in [20]. Though they improved the prediction accuracy over single confidence based approaches [3], [17], the performance is still limited since the combination of confidence features is not optimal.

To select the set of (sub-)optimal confidence features among multiple confidence features, Park and Yoon [24] proposed to utilize the regression forest that computes the importance of confidence features and to train the regression forest classifier using the selected confidence features. Poggi and Mattoccia [25] 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 above methods detect confident pixels at the pixel-level, Kim *et al*. [27] leverages a spatial context to estimate confidence at the superpixel-level by concatenating pixel- and superpixel-level confidence features. However, all of these methods use handcrafted confidence features that might not be optimal to detect unconfident pixels on challenging scenes.

B. CNN-Based Approaches

More recently, several approaches have attempted to estimate the stereo confidence using deep CNNs [28], [29], [42],

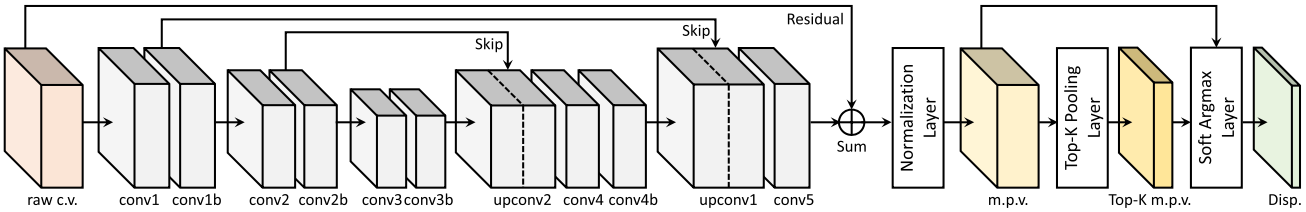


Fig. 2. Illustration of the matching probability construction network.

demonstrating much improved performance over existing handcrafted approaches. In [29], they proposed a novel confidence estimation network that uses a left disparity map only as an input. Seki and Pollefeys [28] proposed to uses both left and right disparity maps in deep networks for improving the confidence prediction accuracy. The confidence refinement network [42] was also developed, which can improve the accuracy of the measured confidence map by leveraging local consistency of confidence map. Shaked and Wolf [32] estimated both disparity and confidence maps from cost volume simultaneously, but they do not use the estimated disparity map as an input of the confidence estimation network.

Tosi *et al.* [43] proposed a novel self-supervised strategy, which generates training labels by leveraging a pool of appropriately combined conventional confidence measures. Poggi *et al.* [44] performed a quantitative evaluation of confidence measures that use machine learning algorithms.

Although CNN-based approaches demonstrated the improved performance compared to handcrafted features-based methods, they use disparity or cost volume only, and no attempt has been made to use both disparity and cost volume. It was shown in existing handcrafted approaches [20], [23], [24], [27] that using the confidence features extracted from the cost volume as well as the disparity improves the confidence prediction accuracy.

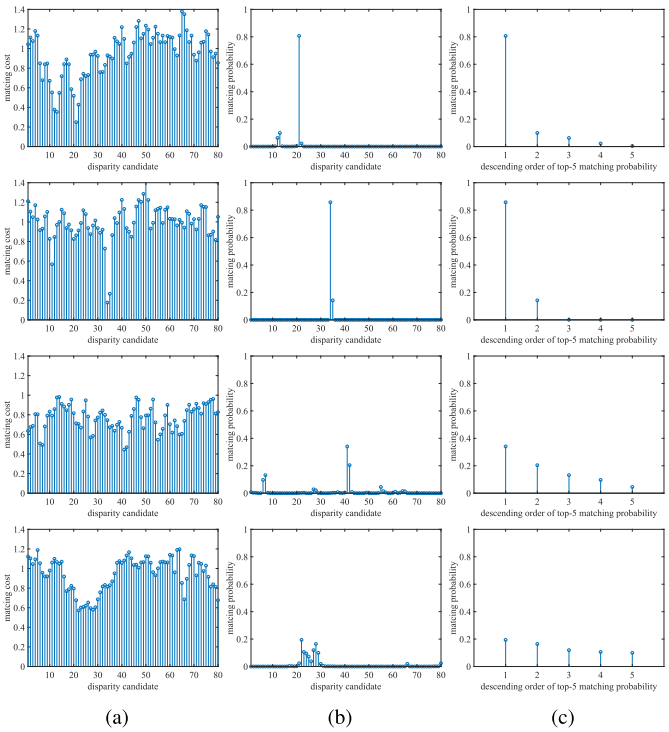


Fig. 3. Effectiveness of top-K matching probability volume: (a) raw matching cost of MC-CNN [9] for the KITTI 2015 dataset [33], (b) matching probability using Eq. (1), and (c) descending order of top-K matching probabilities using Eq. (2) ($K = 5$). First and second row represent reliable pixels, and third and fourth row represent unreliable pixels.

III. Proposed Method

A. Problem Formulation and Model

Let us define a pair of stereo images $\{I^l, I^r\}$. The objective of stereo matching is to estimate a disparity D_i for each pixel $i = [i_x, i_y]^T$ between input stereo images. Since most of stereo matching methods have an estimation error, our approach aims at estimating a confidence of D_i within a learning framework, and refining D_i using the confidence map.

For this task, we propose a novel approach that estimates a confidence map using the matching cost volume and disparity map simultaneously in deep CNNs. The overall network consists of two sub-networks, matching probability construction network (MPCN) and confidence estimation network (CEN). We first estimate a matching cost volume $C_{i,d}$ using existing matching cost functions across a set of disparity candidates $d \in \{0, 1, \ldots, d_{\max}\}$ defined for a maximum disparity $d_{\max}$. Since the direct use of cost volume $C_{i,d}$ to detect the confidence has trouble in solving the scale variation problem and provides a limited discriminative power, we generate the matching probability volume $P_{i,d}$ by refining the initial cost volume $C_{i,d}$ in the matching probability construction network as in Fig. 2. In the confidence estimation network as in Fig. 4, our method then learns the relationship between the matching probability volume $P_{i,d}$ with its associated disparity D_i and a ground truth confidence Q_i^*, computed by thresholding an absolute difference between the estimated disparity D_i and ground truth disparity D_i^*. In this network, the confidence features are extracted using both P_i and D_i. Additionally, motivated by methods using handcrafted confidence features [24], [25], we employ a multi-scale CNN architecture that encodes confidence features from disparity maps at multiple scales. Furthermore, since sparse and insufficient ground truth disparity maps are available in most stereo databases [33], a semi-supervised loss is proposed to learn the network, where we use confident points to compute the image recontruction loss.

B. Confidence Estimation Using Cost Volume and Disparity

1) Matching Probability Construction Network: Although the matching cost volume $C_{i,d}$ includes useful information to find reliably matched pixels, the raw matching cost volume itself has a limited discriminative power and the scale

variation problem, thus direct use of the cost volume provides limited performance for estimating confidences as described in [28] and [44]. To overcome these limitations, we formulate convolutional layers to aggregate and refine the raw matching costs with learned kernels, followed by the normalization layer. After passing the raw matching cost volume through a series of convolutional layers to construct the refined cost volume such that $C' = \mathcal{F}(C; W^c)$ with network parameters W^c, we generate the matching probability volume P, which normalizes the refined matching cost volume C' such that

$$P_{i,d} = \frac{\exp(-C'_{i,d}/\sigma)}{\sum_u \exp(-C'_{i,u}/\sigma)}, \tag{1}$$

where $u \in \{1, \ldots, d_{\max}\}$ and σ is a parameter to adjust a flatness of the matching cost volume. Compared to a softmax module, this normalization scheme can adjust the flatness of matching probability volume according to σ, improving the discriminative power as exemplified in Fig. 3. Note that σ is defined according to the relative scale of the matching cost. As exemplified in Fig. 3(a), the absolute scale of raw matching cost $C_{i,d}$ varies for each pixel, and thus they provide a low discriminative power to be used for estimating reliable pixels. This problem can be alleviated in the matching probability volume $P_{i,d}$ as shown in Fig. 3(b).

Furthermore, we can also find that the majority of matching probabilities $P_{i,d}$ has a value close to 0, which does not convey useful cues as shown in Fig. 3(b). Such redundant parts rather distract the performance of confidence estimation, which will be described in the following section. Thus we also propose a top-K pooling layer, where the matching probability P_i of Eq. (1) is projected into a fixed length input as follows:

$$P_{i,k} = \max_d^k P_{i,d}, \tag{2}$$

where $\max^k(\cdot)$ is the k-th maximal value for $k = \{1, \ldots, K\}$. As shown in Fig. 3(c), top-K matching probabilities have a consistent shape by descending ordering according to reliable or unreliable pixels. Matching outliers in the unreliable pixels lead to rather scattered (or even uniform) distribution of matching probability, while reliable pixels yield concentrated distribution of probability, thus providing a highly discriminative power for confidence estimation. Note that this layer has no trainable parameters, but is differentiable, and thus it can be inserted to any networks as an intermediate layer. In Fig. 3(c), the top-K matching probabilities demonstrate a high discriminative power to classify reliable pixels.

Moreover, since our confidence estimator predicts the confidence using the cost volume and disparity map simultaneously, we additionally compute its associated disparity map. It can be realized through a winner-take-all (WTA) strategy, but it is not differentiable. To overcome this, we use the soft argmin layer as in [45] as follows:

$$D_i = \sum\nolimits_d d \times P_{i,d}. \tag{3}$$

This operation is fully differentiable and allows us to regress the disparity during training. Note that since the matching probability construction network is learned with the estimated confidence simultaneously, which will be described in the following section, the quality of the disparity map D is gradually improved during training. It is used to learn the confidence estimation networks, which enables us to estimate reliable disparity and confidence maps, different from conventional methods that use fixed disparity maps to learn the confidence estimation networks [25], [28], [29], [32].

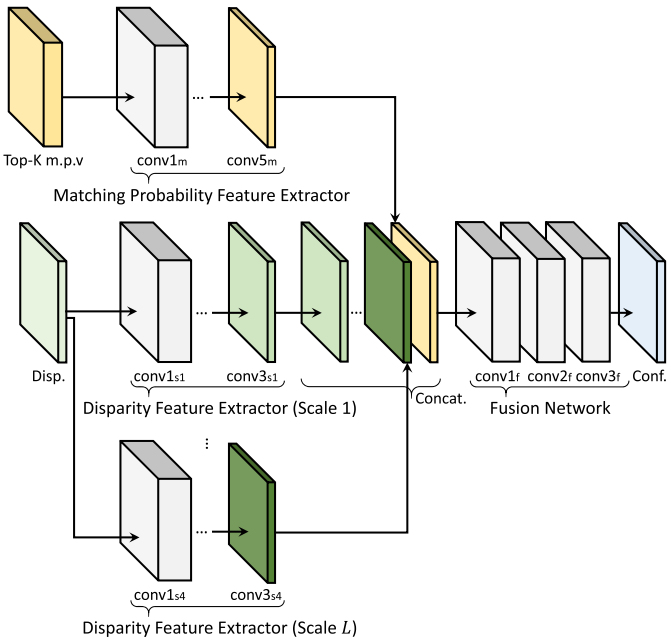


Fig. 4. Illustration of the confidence estimation network.

2) Confidence Estimation Network: At the heart of our strategy to boost the performance of confidence estimation is the fusion scheme of two inputs, consisting of the top-K matching probability volume P and its corresponding disparity map D. Due to heterogeneous attributes of them, a direct concatenation of two raw inputs does not provide an optimal performance. Alternatively, we can simply fuse intermediate outputs obtained from sub-networks for two raw inputs, but the prediction output cannot be fused optimally since the contribution of each input might vary for each pixel.

To overcome these limitations, we build the deep architecture inspired by [46], including a fusion network to estimate the confidence as well as two sub-networks to extract confidence features. Moreover, we extract multi-scale disparity features to improve the discriminative power of confidence features by taking different size of convolutional filters as in [24] and [25]. In order to extract disparity features from different scales, we set different size of convolutional filters for each scale level $l \in \{1, 2, \ldots, L\}$, where L is the number of scales. Each disparity feature extractor network is defined such that $A_l^d = \mathcal{F}(D; W_l^d)$ while matching probability feature extractor network is defined such that $A^c = \mathcal{F}(P; W^c)$, where W_l^d and W^c are network parameters. After all intermediate activations, A^c and A_l^d, are concatenated, the confidence map is finally regressed as the output of fusion network $Q = \mathcal{F}(A^c, A_1^d, \ldots, A_L^d; W^f)$, where W^f is a fusion network parameter and the sigmoid function is used for the last activation to train the binary classifier. Fig. 5 represents the effectiveness of the proposed fusion strategy.

TABLE I
OUR NETWORK ARCHITECTURE OF MATCHING PROBABILITY CONSTRUCTION NETWORK, WHERE 'D.F. I/O' DENOTES DOWNSCALING FACTOR OF INPUT AND OUTPUT FOR EACH LAYER RELATIVE TO THE INPUT IMAGE

Layer	Kernel	Ch I/O	D.F. I/O	Input
conv1	3×3	$d_{\max}$/64	1/1	raw c.v.
conv1b	3×3	64/64	1/1	conv1
pool1	2×2	64/64	1/2	conv1b
conv2	3×3	64/128	2/2	pool1
conv2b	3×3	128/128	2/2	conv2
pool2	2×2	128/128	2/4	conv2b
conv3	3×3	128/128	4/4	pool2
conv3b	3×3	128/128	4/4	conv3
upconv2	4×4	128/128	4/2	conv3b
conv4	3×3	256/128	2/2	upconv2+conv2b
conv4b	3×3	128/64	2/2	conv4
upconv1	4×4	64/64	2/1	conv4b
conv5	3×3	128/$d_{\max}$	1/1	upconv1+conv1b
Sum	1×1	$d_{\max}$/$d_{\max}$	1/1	raw c.v.+conv5
Norm.	-	$d_{\max}$/$d_{\max}$	1/1	Sum
Top-K Pooling	-	$d_{\max}$/K	1/1	Norm.
Soft Argmax	-	$d_{\max}$/1	1/1	Norm.

C. Network Configuration

The deep network for the matching probability construction is illustrated in Fig. 2, and the configuration of this network is summarized in Table I. The matching probability construction network consists of a cost volume refinement network, normalization layer, and top-K pooling layer. To refine the raw matching cost volume, we design the encoder-decoder network with skip layers, consisting of sequential convolutional layers followed by batch normalization (BN) and rectified linear units (ReLU). We apply 2×2 max-pooling operators sequentially, resulting in a total down-sampling factor of 4. In the decoding procedure, we upsample the intermediate features using bilinear deconvolutional filters, and concatenate the upsampled features using a skip layer. Instead of directly predicting the refined cost volume, we predict a residual cost volume by adding a skip connection from the raw cost volume to the output. As the residual learning alleviates the need for restoring specific cost volume contents, our cost volume refinement network provides significant improvement in refining the raw cost volume.

The deep networks for the confidence estimation are shown in Fig. 4, and the configuration of this network is summarized in Table II. For the confidence estimation network, we leverage 3×3 convolutional filters sequentially, followed by BN and ReLU. In this network, the pooling operator is not used to preserve the spatial resolution. To extract multi-scale disparity features with different sizes of receptive fields, we vary the convolutional filter size relative to scale level $l \in \{1, 2, \ldots, L\}$ as explained in Table II. Here, we set L as 4. We also apply Sigmoid operator on the output of the last activation since the output confidence value is within $[0, 1]$.

D. Loss Function

1) Matching Probability Construction Network Loss: A major challenge of CNN-based stereo matching methods is the lack of dense ground truth disparity maps especially in outdoor settings [47]. To overcome this limitation, we propose a semi-supervised learning scheme in a manner that highly confident pixels in the disparity map are used together with sparse ground truth disparities to learn the matching probability construction network. Specifically, with the sparse ground truth disparity D^* and the estimated disparity D, we define a novel loss function for the MPCN, consisting of a supervised term $\mathcal{L}_{sup}$ and an unsupervised term $\mathcal{L}_{unsup}$ with hyper-parameter λ that weighs the contribution of the latter:

$$\mathcal{L}_{MPCN} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{unsup}. \tag{4}$$

The supervised term $\mathcal{L}_{sup}$ is formulated to directly regress the disparity D with respect to ground truth disparity D^* with L_1 norm objective function such that

$$\mathcal{L}_{sup} = \sum\nolimits_i m_i \left| D_i - D_i^* \right|, \tag{5}$$

where m_i is a mask that represents the existence of ground truth disparity map. It should be noted that this supervised loss $\mathcal{L}_{sup}$ can provide reliable performance when the network is learned with a sufficient amounts of training data. To overcome this, we additionally propose an unsupervised term $\mathcal{L}_{unsup}$ that leverages the estimated confidence as follows:

$$\mathcal{L}_{unsup} = \sum\nolimits_i \mathcal{G}(Q_i; \rho) \left| I_{i-D_i}^l - I_i^r \right|, \tag{6}$$

where $\mathcal{G}(Q_i; \rho)$ is a truncation function which has the value 1 when the estimated confidence Q_i is higher than a threshold parameter ρ, and 0 otherwise. As ρ decreases, the density of regions used to compute the loss $\left| I_{i-D_i}^l - I_i^r \right|$ increases, but it is more likely that the convergence of networks is disturbed due to unreliable regions used in the loss function. Contrarily, as ρ increases, the density of regions used to compute the loss decreases, slowing down the convergence. Here, we set ρ in which the density of confident regions used in the loss computation becomes 75% on average. This loss function is similar to the image recostruction loss as in [36] that minimizes the difference between right color image I_i^r and warped left color $I_{i-D_i}^l$. One difference is that our loss function only considers highly confident regions to learn the disparity. To achieve a convergent optimization and avoid the bias problem, we first learn the confidence map with only the supervised loss $\mathcal{L}_{sup}$ until the pre-defined number of epoch, and then fine-tuned them with the semi-supervised loss (i.e., $\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$). After learning the network, we put the semi-supervised loss in Eq. (6) together with Eq. (5) with reliable confidence. With these loss functions, our method reliably learns the matching probability construction network with the datasets having only sparse ground truth disparity map. Fig. 6 shows the importance of a semi-supervised loss function for learning matching probability construction network.

2) Confidence Estimation Network Loss: Compared to the matching probability construction network, the confidence estimation network needs relatively lower number of parameters, and such shallower networks can be learned even with limited sparse disparities as in [28] and [29]. The loss function of the CEN is designed to predict the confidence Q with respect to the ground truth confidence Q^* defined as follows:

$$Q_i^* = \mathcal{G}(|D_i^* - D_i|; \delta), \tag{7}$$

TABLE II

OUR NETWORK ARCHITECTURE OF CONFIDENCE ESTIMATION NETWORK

Matching Probability Feature Extractor				
Scale	Layer	Kernel	Ch I/O	Input
-	conv1_m	3×3	K/256	top-K m.p.v.
	conv2_m	3×3	256/128	conv1_m
	conv3_m	3×3	128/64	conv2_m
	conv4_m	3×3	64/32	conv3_m
	conv5_m	3×3	32/1	conv4_m
Disparity Feature Extractor				
Scale	Layer	Kernel	Ch I/O	Input
Scale1	conv1_{s1}	3×3	1/128	disparity
	conv2_{s1}	3×3	128/64	conv1_{s1}
	conv3_{s1}	1×1	64/1	conv2_{s1}
Scale2	conv1_{s2}	5×5	1/128	disparity
	conv2_{s2}	3×3	128/64	conv1_{s2}
	conv3_{s2}	1×1	64/1	conv2_{s2}
Scale3	conv1_{s3}	7×7	1/128	disparity
	conv2_{s3}	3×3	128/64	conv1_{s3}
	conv3_{s3}	1×1	64/1	conv2_{s3}
Scale4	conv1_{s4}	9×9	1/128	disparity
	conv2_{s4}	3×3	128/64	conv1_{s4}
	conv3_{s4}	1×1	64/1	conv2_{s4}
Fusion Network				
Scale	Layer	Kernel	Ch I/O	Input
-	conv1_f	3×3	5/64	conv4_m+conv2_{s1}+...+conv2_{s4}
	conv2_f	3×3	64/32	conv1_f
	conv3_f	3×3	32/1	conv2_f
	Sigmoid	1×1	1/1	conv3_f

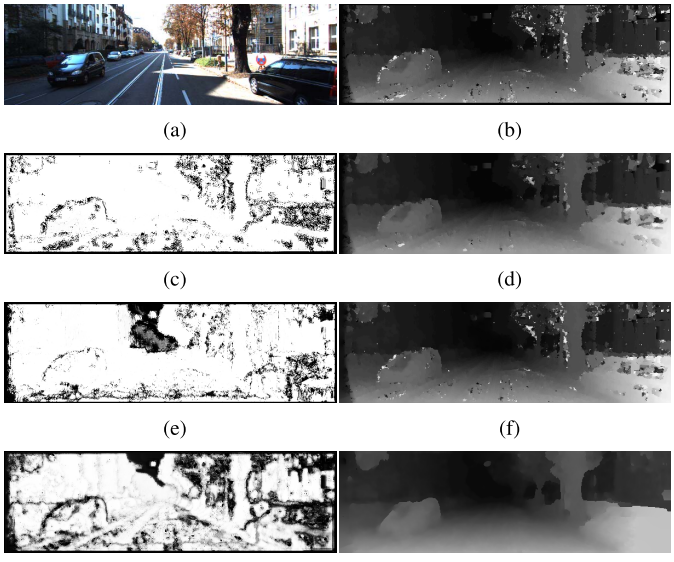


Fig. 5. Importance of a joint use of matching probability volume and disparity map in the confidence estimation: (a) left color image, (b) an initial disparity map estimated using MC-CNN [9]. Confidence maps were estimated using the matching probability volume only in (c), disparity map only in (e), and both matching probability volume and disparity map in (g), respectively. (d), (f), and (h) represent disparity map refined using the confidence maps in (c), (e), and (g), respectively. By jointly using the matching probability volume and disparity map, our method provide reliable confidence estimation performance.

where δ is a threshold parameter to compare the ground truth disparity D^* and the estimated disparity D which is the output of the matching probability construction network.

The loss function for the confidence estimation network $\mathcal{L}_{CEN}$ is then defined as the cross-entropy loss:

$$\mathcal{L}_{CEN} = -\sum_i \left[Q_i^* \log Q_i + (1 - Q_i^*) \log(1 - Q_i) \right]. \quad (8)$$

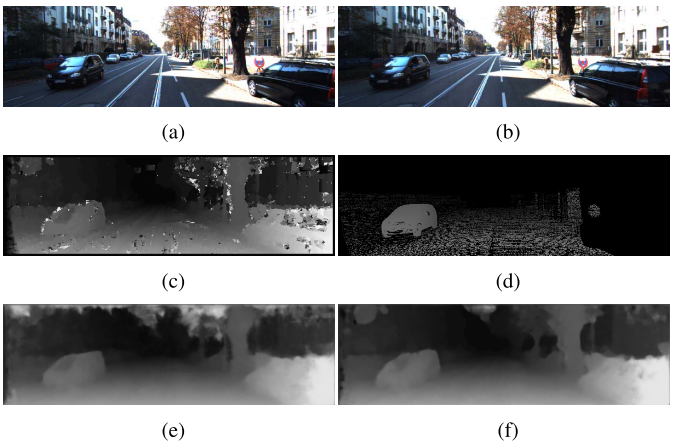


Fig. 6. Importance of a semi-supervised loss for learning the matching probability construction network: (a) left and (b) right color images, (c) an initial disparity map estimated using MC-CNN [9], (d) sparse ground truth disparity map. Disparity maps estimated from matching probability construction network learned (e) with only supervised loss ($\mathcal{L}_{sup}$) and (f) with semi-supervised loss ($\mathcal{L}_{sup}$ and $\mathcal{L}_{unsup}$). The errors within the regions where the ground truth is not defined are efficiently removed with unsupervised loss.

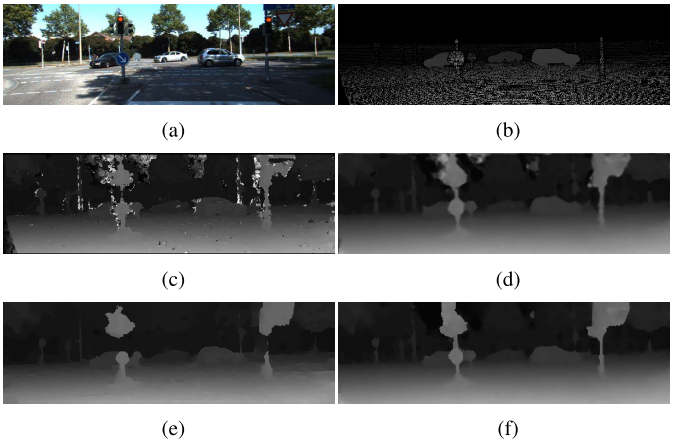


Fig. 7. The effectiveness of the proposed matching probability volume network: (a) left color image, (b) sparse ground truth disparity map, (c) an initial disparity map estimated using MC-CNN [9], (d) an intermediate disparity map obtained by the proposed matching probability volume network. (e) and (f) represent refined disparity map of (c) and (d), respectively.

Note that the ground truth confidence Q^* varies during training unlike existing confidence estimation approaches in [28] and [29] which Q^* is fixed during training. Our key contribution is to refine the cost volume during training to provide highly discriminative confidence features and refined disparity maps. Note that Q^* is computed by comparing the resultant disparity map and the ground truth disparity map. Namely, the cost volume refinement (CVR) network evolves and improves the matching cost and its associated disparity map during training, leading to varying ground truth confidence Q^*.

IV. VALIDATION

So far we have explained the CNN-based approach to improve the accuracy of the confidence and disparity maps in a boosting manner. By evolving the training, the quality of intermediate disparity maps is improved implicitly in Fig. 7(d) when compared to an initial disparity map in Fig. 7(c).

TABLE III
EXPERIMENTAL CONFIGURATION

Raw cost	Census-SGM / MC-CNN		
Training Set	MPI Sintel		KITTI 2012 (w/GT) + KITTI raw data (wo/GT)
Testing Set	MID 2006	MID 2014	KITTI 2015

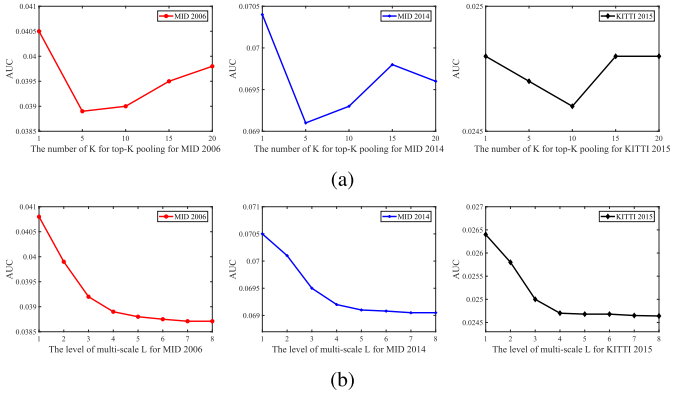


Fig. 8. Ablation study: (a) Average AUC value of different K for top-K pooling layer and (b) different level of multi-scale L for MID 2006, MID 2014, and KITTI 2015 dataset (from left to right). The raw matching cost is MC-CNN [9].

However, the performance gain is still limited as no confidence map is involved in the cost volume refinement step [32]. In this section, we introduce three validatation methods to prove the effectiveness of our confidence measure in the post-processing step of the stereo matching pipeline. With the post-processing steps, the erroneous disparities are corrected using the estimated confidence map as shown in Fig. 7(f).

A. Cost Modulation Based Optimization

Following [24], we first incorporate the predicted confidence map into stereo matching algorithms by modulating the refined cost volume such that

$$\widehat{C}_{i,d} = Q_i C'_{i,d} + (1 - Q_i) \sum_d C'_{i,d} / d_{\max}. \tag{9}$$

After modulating the refined cost volume, the refined cost of confident pixels remain unchanged while those of unconfident pixels are flattened. Thus unconfident pixels can be easily dominated by confident neighboring pixels in the optimization step. To produce a final disparity map $\widehat{D}$, the modulated cost function $\widehat{C}$ is then optimized using global approaches such as SGM [8] or belief propagation [48].

B. GCPs-Based Optimization

The predicted confidence can also be incorporated in GCPs-based optimization as in [38]. We first set reliable pixels that have a higher value than a threshold τ as the GCPs, and then globally propagate the initial GCPs through an MRF-based optimization by minimizing the following energy function:

$$E_{GCP} = \sum_i \left(h_i (\widehat{D}_i - D_i)^2 + \lambda_1 \sum_{j \in \mathcal{N}_i^4} w_{i,j}^I (\widehat{D}_i - \widehat{D}_j)^2 \right), \tag{10}$$

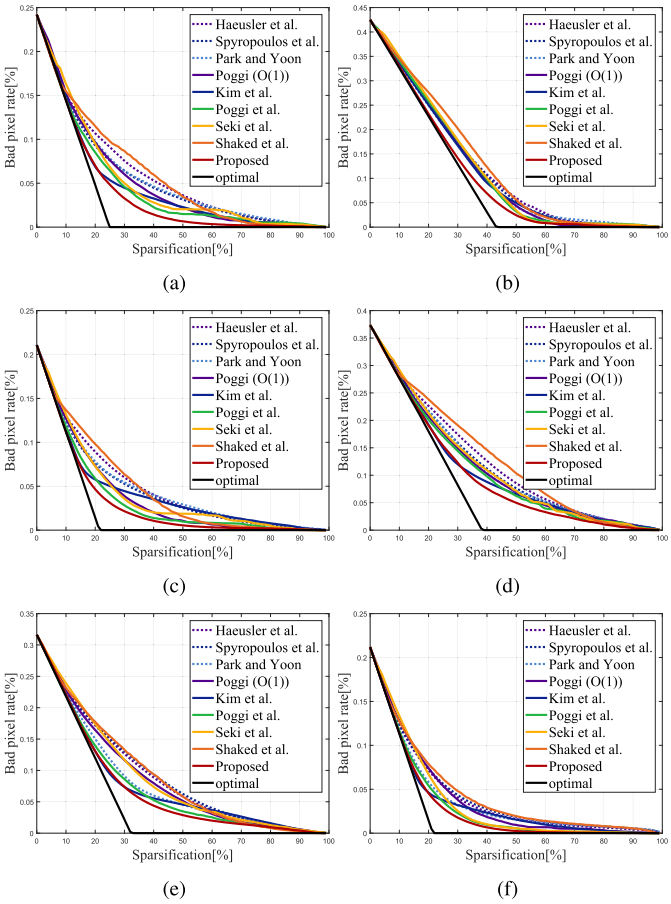


Fig. 9. Sparsification curve for (a) 'Baby3', (b) 'Midd2', and (c) 'Wood1' images selected from the MID 2006 dataset [39] using census-SGM, and (d) 'Mask', (e) 'Playtable', and (f) 'Sticks' images selected from the MID 2014 dataset [40] using MC-CNN. The sparsification curve for a ground truth confidence map is described as 'optimal'.

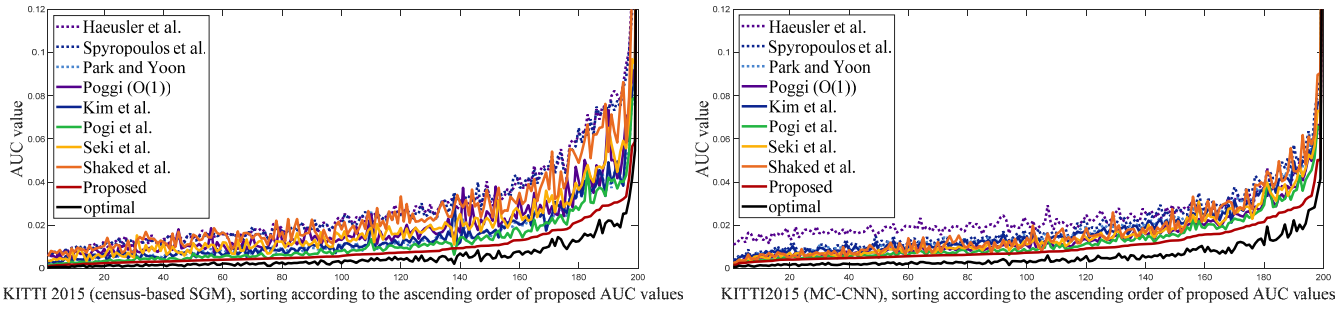
where h_i is the binary mask to indicate the GCPs, and $\widehat{D}_i$ is output disparity map. $w_{i,j}^I$ is the affinity between i and j in the feature space consisting of color I and spatial location, and $\mathcal{N}_i^4$ represents a local 4-neighborhood. This simple quadratic optimization can be efficiently solved using [38].

C. Aggregated GCPs-Based Optimization

We further utilize an aggregated data term to mitigate propagation errors by inaccurately estimated confidences when interpolating a sparse disparity map obtained by thresholding the confidence map. It was shown in [38] that a more robust data constraint using an aggregated data term leads to a better quality in the sparse data interpolation. In this regard, we define the energy function as follows:

$$E_{AGCP} = \sum_i \left(\sum_{r \in \mathcal{M}_i} h_r c_{i,r}^I (\widehat{D}_i - D_r)^2 + \lambda_2 \sum_{j \in \mathcal{N}_i^4} w_{i,j}^I (\widehat{D}_i - \widehat{D}_j)^2 \right), \tag{11}$$

where $\mathcal{M}_i$ represents a set of neighborhoods used to aggregate the disparity map filtered out with the confidence Q. Note that



(a) (b)

Fig. 10. AUC values of (a) census-based SGM and (b) MC-CNN for the KITTI 2015 dataset [33]. We sort the AUC values in the ascending order according to the proposed AUC values. The 'optimal' AUC values are calculated using ground truth confidence map.

TABLE IV

THE AVERAGE AUC VALUES FOR MID 2006 [39], MID 2014 [40], AND KITTI 2015 [33] DATASET. THE AUC VALUE OF GROUND TRUTH CONFIDENCE IS MEASURED AS 'OPTIMAL'. THE RESULT WITH THE LOWEST AUC VALUE IN EACH EXPERIMENT IS HIGHLIGHTED

Dataset	MID 2006 [39]		MID 2014 [40]		KITTI 2015 [33]	
Raw cost	Census-SGM	MC-CNN	Census-SGM	MC-CNN	Census-SGM	MC-CNN
Haeusler et al. [23]	0.0454	0.0417	0.0841	0.0750	0.0585	0.0308
Spyropoulos et al. [20]	0.0447	0.0420	0.0839	0.0752	0.0536	0.0323
Park and Yoon [24]	0.0438	0.0426	0.0802	0.0734	0.0527	0.0303
Poggi et al. (O(1)) [25]	0.0439	0.0413	0.0791	0.0707	0.0461	0.0263
Kim et al. [27]	0.0430	0.0409	0.0772	0.0701	0.0430	0.0294
Poggi et al. [29]	0.0454	0.0402	0.0769	0.0716	0.0419	0.0258
Seki et al. [28]	0.0462	0.0413	0.0791	0.0718	0.0439	0.0272
Shaked et al. [32]	0.0464	0.0495	0.0806	0.0736	0.0531	0.0292
Proposed wo/CVR [41]	0.0451	0.0401	0.0758	0.0701	0.0412	0.0254
Proposed wo/Top-K	0.0420	0.0395	0.0748	0.0700	0.0408	0.0257
Proposed wo/MDF	0.0423	0.0392	0.0741	0.0697	0.0411	0.0252
Proposed	**0.0417**	**0.0389**	**0.0730**	**0.0692**	**0.0403**	**0.0247**
Optimal	0.0340	0.0323	0.0569	0.0527	0.0348	0.0170

$\mathcal{M}_i$ is not limited to $\mathcal{N}_4$ neighbors, but usually more neighbors are used for ensuring a large support according to [38]. We define $c^I_{i,r}$ using a bilateral kernel between pixel i and r in the feature space consisting of color intensity I, spatial location, and estimated confidence as used in [27]. $w^I_{i,j}$ is defined similar to the above section. This optimization can also be efficiently solved using constant-time edge-aware filtering and sparse matrix solver [38].

V. EXPERIMENTAL RESULTS

A. Experimental Settings

In the following, we evaluated the proposed method compared to conventional shallow classifier based approaches such as Haeusler *et al.* [23], Spyropoulos *et al.* [20], Park and Yoon [24], Poggi and Mattoccia (O(1)) [25], Kim *et al.* [27], and CNN-based approaches such as Poggi and Mattoccia [29], Seki and Pollefey [28], and Shaked and Wolf [32]. We obtained the results of [24] by using the author-provided MATLAB code, while the results of [20], [23], [28], and [32] were obtained using our own MATLAB implementation. For [25] and [29], we re-impelemented the algorithms based on author's lua code. For an evaluation, we used the Middlebury (MID) 2006 [39], MID 2014 [40], and MPI [34] dataset taken or synthesized under carefully-controlled environments, and real-world datasets as KITTI [33], as described in Table III. For MID 2006 and MID 2014, we trained the classifier using 981 images in the MPI dataset and for KITTI 2015 dataset, we trained the classifier using 194 stereo pairs in KITTI 2012 dataset [33] for supervised loss $\mathcal{L}_{sup}$, and 40,000 raw stereo image pairs for unsupervised loss $\mathcal{L}_{unsup}$. To evaluate the performance of the confidence measure quantitatively, we used the sparsification curve and its area under curve (AUC) as in [20], [23], [24], and [28]. The sparsification 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 prediction errors. AUC quantifies the ability of a confidence measure to estimate correct matches. The higher the accuracy of the confidence measure is, the lower the AUC value is. To evaluate the disparity refinement performance using the estimated confidence map, we also measured the average bad matching percentage (BMP) as in [39].

B. Implementation Details

Our networks are trained in an end-to-end manner, given the raw cost volume as an input and the ground truth disparity and ground truth confidence as outputs. For training and testing networks, we used the `VLFeat MatConvNet`

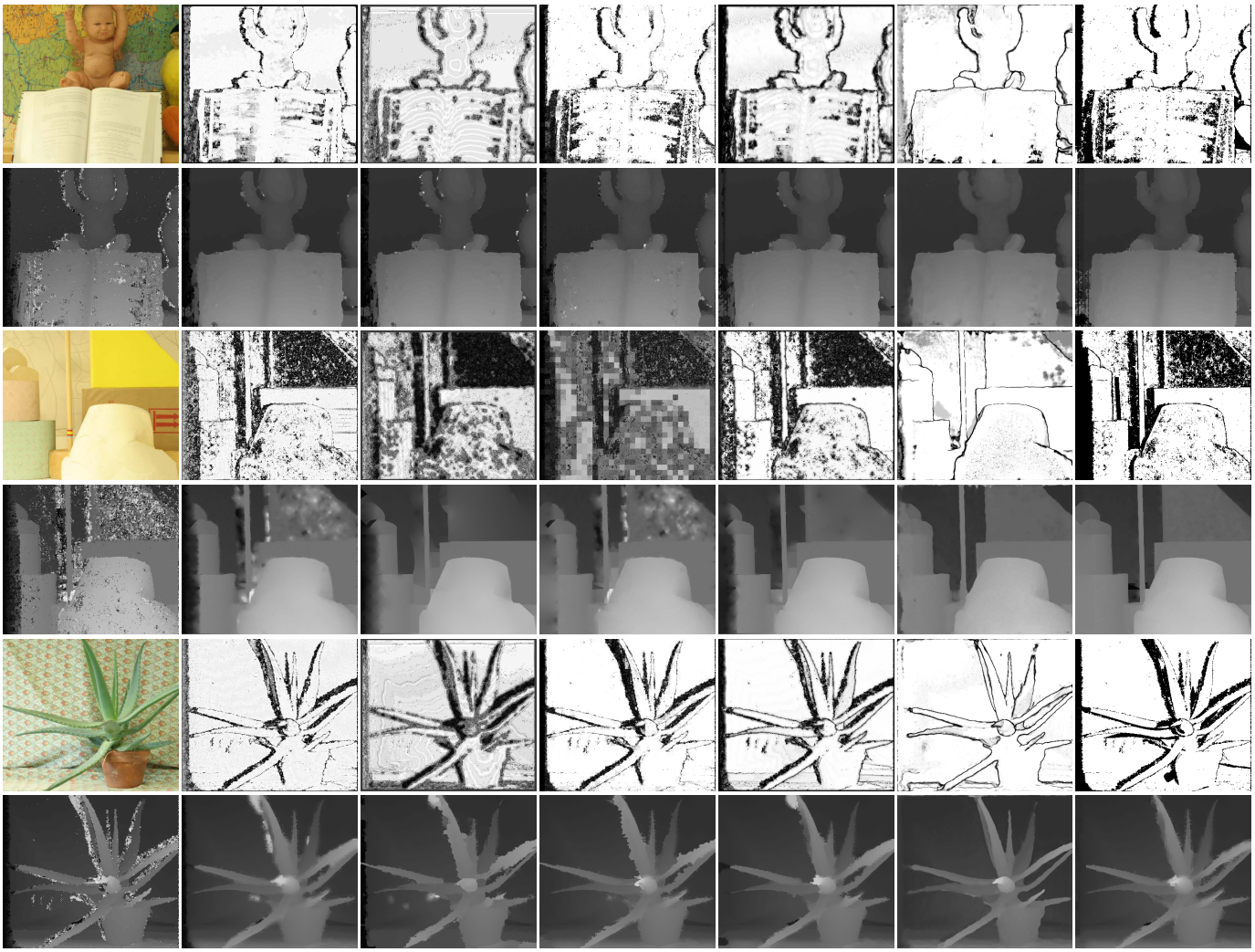


Fig. 11. The confidence and refined disparity maps on the MID 2006 dataset [39] using census-based SGM+mod. (1st and 2nd rows), +GCPs (3rd and 4th rows), and +AGCPs (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Park and Yoon [24], (c) Poggi *et al.* (O(1)) [25], (d) Kim *et al.* [27], (e) proposed wo/CVR [41], (f) proposed, and (g) ground truth confidence map.

`toolbox` [49]. We make use of the stochastic gradient descent with momentum, and set the learning rate $1e-4$ and the batch size 50. The proposed method was implemented in MATLAB and was simulated on a PC with TitanX GPU. To compute a raw matching cost, we used a census transform with a 5×5 local window followed by SGM and the matching cost with a convolutional neural network (MC-CNN) method [9]. For computing the MC-CNN, 'KITTI 2012 fast network' was used, provided at the author's website [50]. For the census transform, we applied SGM [8] on the estimated cost volume by setting $P_1 = 0.008$ and $P_2 = 0.126$ as in [24]. We set ρ to 0.9 at which the density of confident regions used in the loss computation becomes 75% on average. The flatness parameter σ for obtaining the matching probability volume is set as 100 and 0.05 for SGM and MC-CNN respectively, according to their relative scales. The pre-defined number of epoch was set empirically to 200 at which the networks is pre-trained and converged with the supervised loss only. The hyper-parameter λ is set to 1.

C. Component Analysis

We first analyzed the performance of the proposed confidence estimation method as varying the number of K in top-K pooling and the scale L for multi-scale disparity feature extraction in confidence estimation network. AUC values as varying K are shown in Fig. 8(a). If K is too small, the discriminative power becomes low. As K increases over 5, the AUC value is degraded because of the redundancy. We set K as 5 in all experiments. Fig. 8(b) shows the convergence anlysis for varying numbers of scale L from 1 to 8. The AUC value decreases as L increases, and converges after $L = 4$. Based on these experiments, we set $L = 4$ for considering the trade-off between accuracy and complexity.

Moreover, for ablation experiments to validate the componets, we evaluated the proposed method without cost volume refinement network (Proposed wo/CVR) [41] which is the previous version of the proposed method, without top-K pooling layer (Proposed wo/Top-K), and without multi-scale disparity

Fig. 12. The confidence and refined disparity maps on MID 2006 dataset [39] using MC-CNN+mod. (1st and 2nd rows), +GCPs (3rd and 4th rows), and +AGCPs (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Haeusler *et al*. [23], (c) Poggi *et al*. [29], (d) Seki and Pollefeys [28], (e) proposed wo/CVR [41], (f) proposed, and (g) ground truth confidence map.

feature extraction (Proposed wo/MDF) in quantitative results of average AUC values for various datasets as in Table IV. Quantitative result also shows the effectiveness of multi-scale disparity feature extraction (Proposed wo/MDF) and top-K pooling layer (Proposed wo/Top-K).

D. Confidence Measure Analysis

We compared the AUC of our method with conventional learning-based approaches using handcrafted confidence features [20], [23]–[25], [27] and CNN-based methods [28], [29], [32]. The optimal AUC can be obtained with a ground truth confidence map. Sparsification curves for selected frames in the MID 2006 [39] and MID 2014 dataset [40] with census-based SGM and MC-CNN are shown in Fig. 9. The results have shown that the proposed confidence estimator exhibits a better performance than conventional handcrafted approaches [20], [23]–[25], [27] and CNN-based approaches [28], [29], [32]. Fig. 10 describes the AUC values, which are sorted in ascending order, for the KITTI 2015 dataset [33] with census-based SGM and MC-CNN respectively. The handcrafted approaches showed inferior performance than the proposed method due to low discriminative power of the handcrafted confidence features. CNN-based methods [28], [29], [32] have improved confidence estimation performance compared to existing handcrafted methods such as [24], but they are still inferior to our method as they rely on either used only estimated disparity maps or cost volume to predict unreliable pixels.

The estimated confidence maps are shown in Fig. 11 - Fig. 16. The resultant confidence maps of our method looks different since the confidence maps of ours were obtained from the matching probability volume while others are obtained from the raw matching cost volume. Note that the confidence map can also be predicted directly from the matching cost volume without the CVR network, named as 'Proposed wo/CVR [41]' in Fig. 11 - Fig. 14. Experimental results demonstrate that the CVR network enables us to estimate more accurate confidence and disparity maps simultaneously in a boosting manner. Moreover, the average AUC value with census-based SGM and

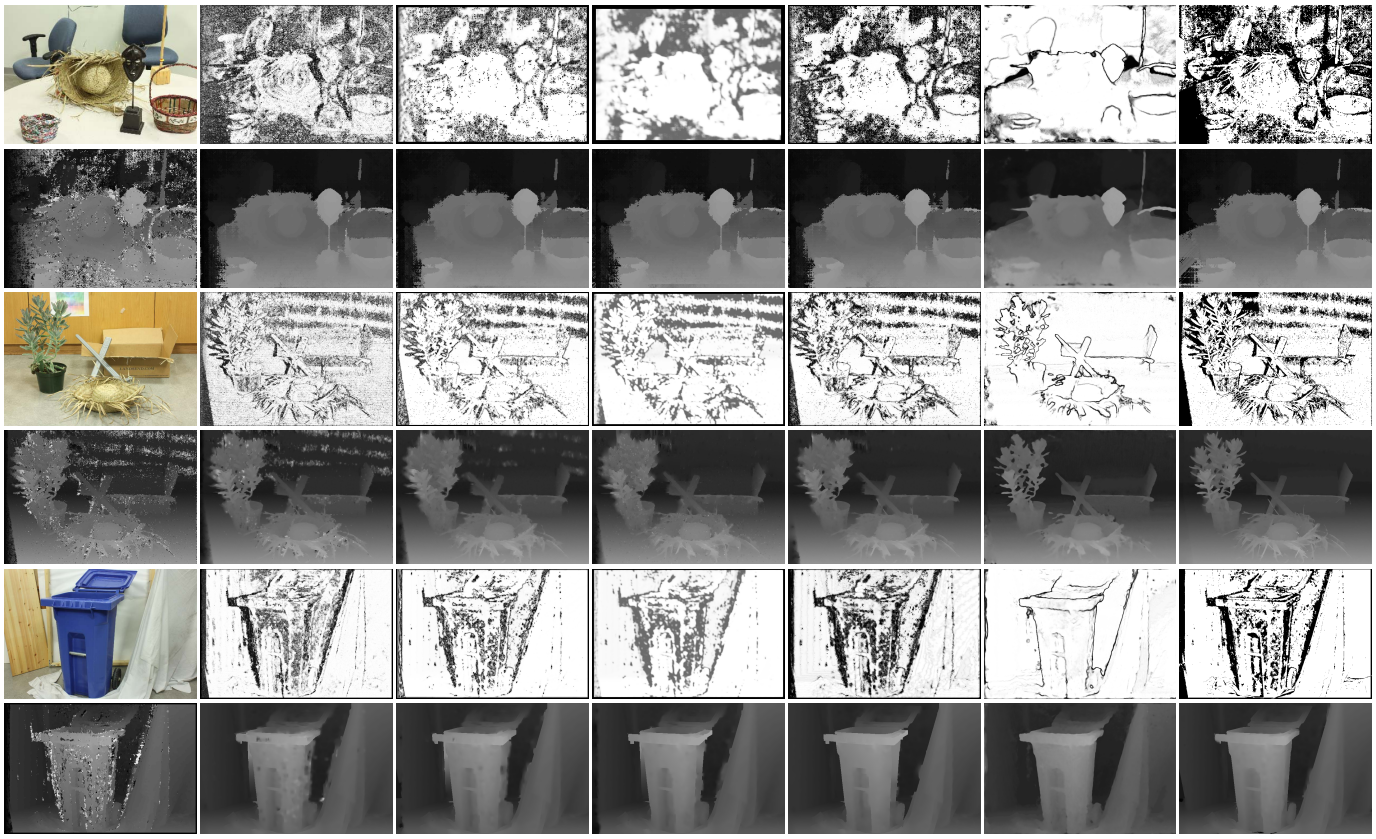


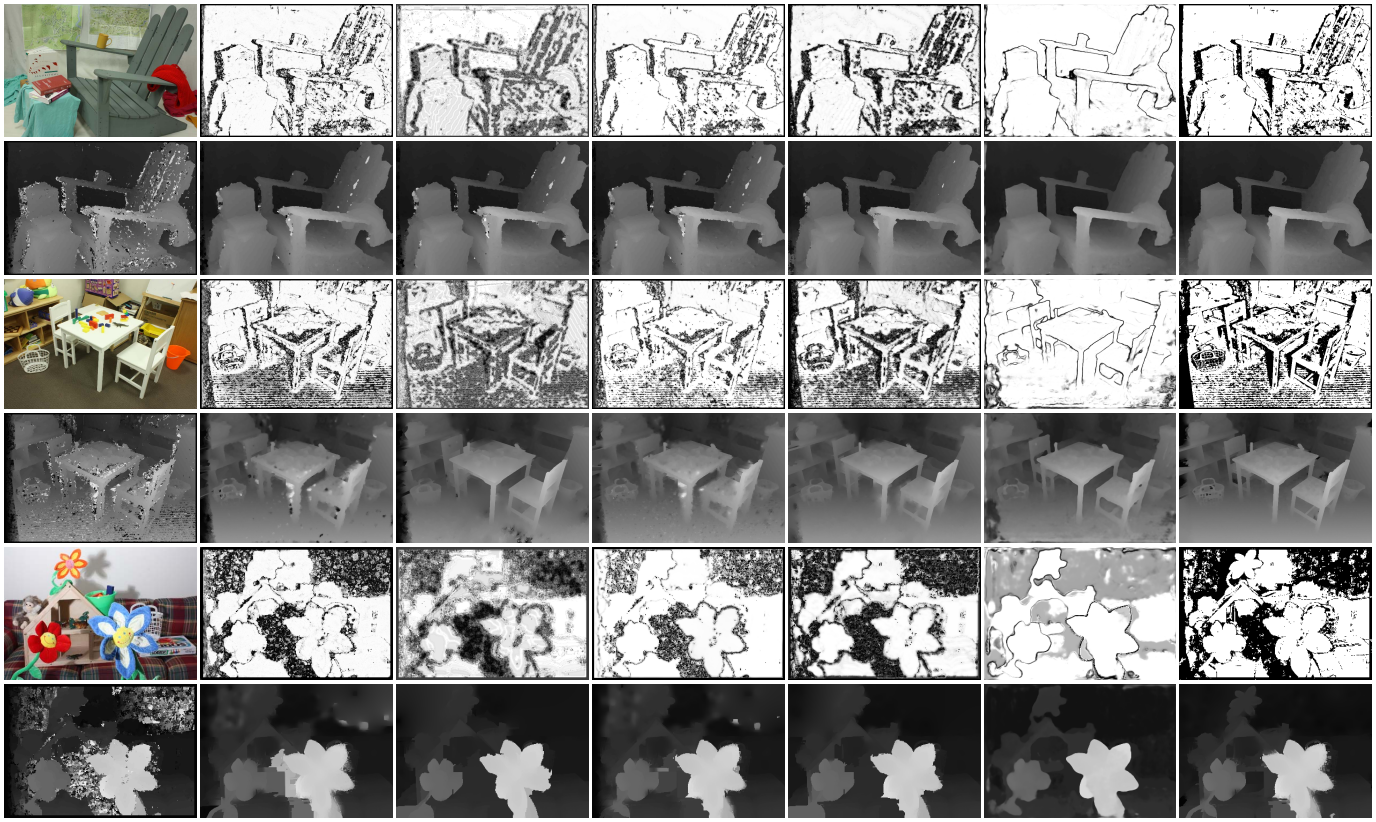
Fig. 13. The confidence and refined disparity maps on MID 2014 dataset [40] using census-based SGM+mod. (1st and 2nd rows), +GCPs (3rd and 4th rows), and +AGCPs (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Spyropoulos *et al.* [20], (c) Poggi *et al.* [29], (d) Seki and Pollefeys [28], (e) proposed wo/CVR [41], (f) proposed, and (g) ground truth confidence map.

MC-CNN for MID 2006, MID 2014, KITTI 2015 datasets were summarized in Table IV. The proposed method always yield the lowest AUC values, showing the superiority of the proposed method compared to the existing CNN-based classifiers [28], [29], [32].

E. Stereo Matching Analysis

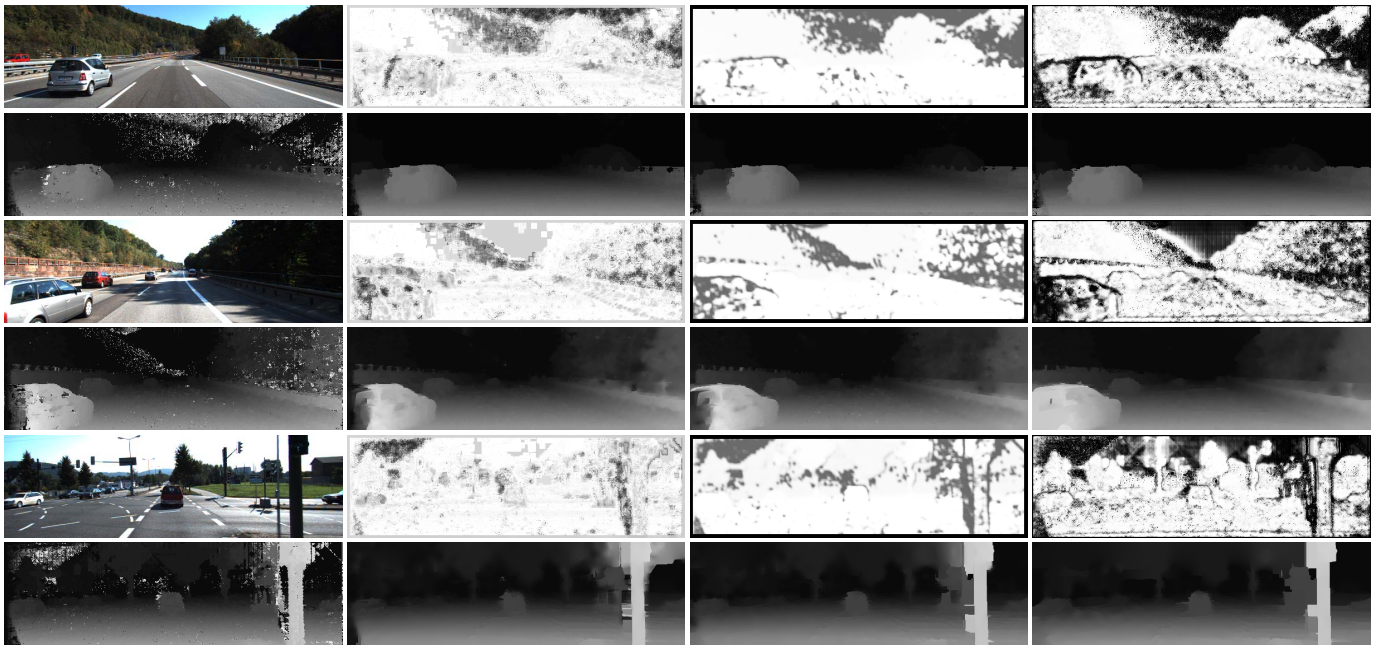
To verify the robustness of the confidence measures, we refined the disparity map using the confidence maps estimated by several confidence measure approaches including ours. For refining the disparity maps, we used three different schemes described in Sec. 4, which are cost modulation (mod.) based optimization [24] and GCPs-based optimization (GCPs) [37], and aggregated GCPs-based optimization (AGCPs) [38] without additional post-processing to clearly show the performance gain achieved by the confidence measure. Note that there exist several literatures to evaluate the confidence measures, but this experiment is the first attempt to evaluate various confidence measures including both conventional methods and CNN-based methods in refining the disparity map. To evaluate the quantitative performance, we measured an average BMP for the MID 2006 [39], MID 2014 [40], and KITTI 2015 [33] datasets. Table V and Table VI show the BMP with one/three pixels when using census-based SGM and MC-CNN respectively. For MID 2006 and MID 2014, since there are occluded pixels in ground truth disparity map, we computed the BMP only for visible pixels. The KITTI 2015 benchmark provides a sparse ground truth disparity map thus we evaluated the BMP only for sparse pixels with the ground truth disparity values. Note that the optimal percentage of BMP was obtained by measuring the ratio of erroneous pixels in which the absolute difference between the disparity map refined using ground truth confidence map and the ground truth disparity map is larger than one or three pixels, respectively. The proposed method achieves the lowest BMP in all experiments.

Fig. 11 - Fig. 16 display the disparity maps refined with the confidence maps estimated from the existing handcrafted classifiers [20], [24], [25], [27] CNN-based classifiers [29], [28], [32] and the proposed method. SGM modulation, GCPs-based optimization, and aggregated GCPs-based optimization were used to refine the disparity maps for the MID 2006 [39], MID 2014 [40], and KITTI 2015 [33] datasets with census-based SGM and MC-CNN respectively. It was clearly shown that the erroneous matches are reliably removed using the proposed confidence measure. For the KITTI 2015 dataset [33], erroneous disparities usually occur in textureless regions (sky and road) as shown in Fig. 15 and Fig. 16. Conventional approaches [25], [28], [29] show the limited performance for detecting incorrect pixels in



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

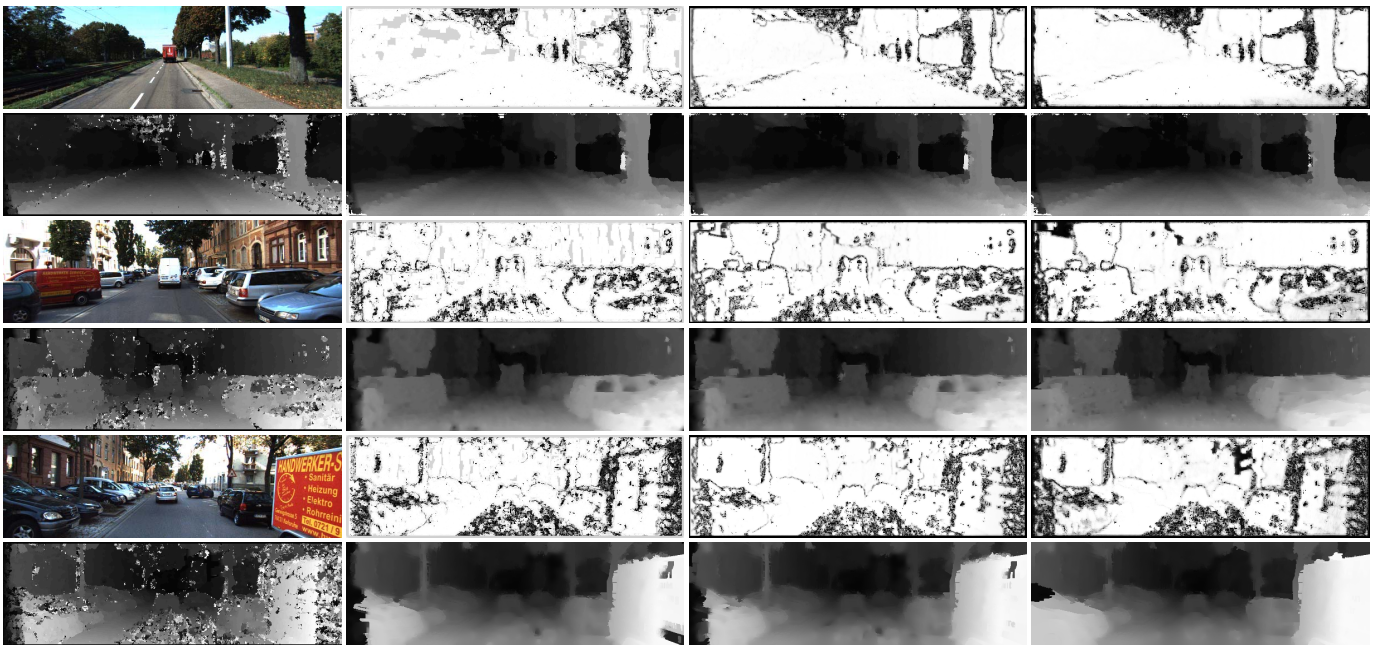
Fig. 14. The confidence and refined disparity maps on MID 2014 dataset [40] using MC-CNN+mod. (1st and 2nd rows), +GCPs.(3rd and 4th rows), and +AGCPs. (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Park and Yoon [24], (c) O(1) [25], (d) Poggi *et al.* [29], (e) proposed wo/CVR [41], (f) proposed, and (g) ground truth confidence map.



(a) (b) (c) (d)

Fig. 15. The confidence and refined disparity maps on KITTI 2015 dataset [33] using census-based SGM+mod. (1st and 2nd rows), +GCPs (3rd and 4th rows), and +AGCPs (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Poggi *et al.* O(1) [25], (c) Seki and Pollefeys [28], and (d) ours.

textureless regions, and thus they affect the matching quality of the subsequent disparity estimation pipeline. In contrast, the proposed method can detect mismatched pixels more reliably.



(a) (b) (c) (d)

Fig. 16. The confidence and refined disparity maps on KITTI 2015 dataset [33] using MC-CNN+mod. (1st and 2nd rows), +GCPs (3rd and 4th rows), and +AGCPs (5th and 6th rows) (a) Color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Poggi *et al.* O(1) [25], (c) Poggi *et al.* [29], and (d) ours.

TABLE V

THE BMP OF THE RESULTANT DISPARITY MAP ON MID 2006 [39], MID 2014 [40], AND KITTI 2015 [33] DATASET WITH CENSUS-BASED SGM. THE BAD PIXEL ERROR RATE OF THE REFINED DISPARITY MAP USING GROUND TRUTH CONFIDENCE IS MEASURED AS 'OPTIMAL'. THE BMP IS MEASURED WITH ONE/THREE PIXEL ERRORS. THE RESULT WITH THE LOWEST BMP IN EACH EXPERIMENT IS HIGHLIGHTED

Dataset	MID 2006 [39]			MID 2014 [40]			KITTI 2015 [33]		
Validations	+mod.	+GCPS.	+AGCPs.	+mod.	+GCPS.	+AGCPs.	+mod.	+GCPS.	+AGCPs.
Initial disparity	20.43/10.18	20.43/10.18	20.43/10.18	43.17/24.46	43.17/24.46	43.17/24.46	23.50/18.67	23.50/18.67	23.50/18.67
Haeusler et al. [23]	16.85/8.36	12.66/7.17	11.82/7.06	38.64/16.55	35.45/12.83	33.66/10.81	17.51/15.19	13.71/11.82	11.56/10.18
Spyropoulos et al. [20]	15.73/7.94	12.44/6.89	11.36/6.51	36.19/16.35	32.29/12.26	31.97/10.62	17.08/14.81	13.18/11.36	11.46/9.59
Park and Yoon [24]	14.40/7.12	11.22/6.35	10.98/6.03	35.83/15.75	31.08/11.79	29.16/10.15	16.92/14.02	12.57/9.76	10.37/9.18
Poggi et al. (O(1)) [25]	14.46/6.85	10.80/5.94	9.58/5.77	34.17/14.37	30.52/10.06	27.03/9.90	16.22/13.51	10.22/9.08	9.69/8.12
Kim et al. [27]	13.77/6.13	8.77/5.17	7.72/5.08	32.09/14.96	28.76/9.75	23.74/8.53	14.97/12.68	9.86/8.43	8.50/7.08
Poggi et al. [29]	13.82/6.14	8.62/4.86	7.84/4.63	31.59/13.51	28.71/9.39	23.98/8.89	14.22/12.13	9.91/8.11	8.68/6.79
Seki et al. [28]	13.39/5.96	7.37/4.57	7.62/4.02	31.13/13.34	27.13/8.75	23.16/8.88	14.48/11.96	8.85/7.03	7.56/6.46
Shaked et al. [32]	16.43/8.06	11.77/7.21	10.42/7.18	37.98/16.52	34.11/12.07	32.08/11.41	17.88/15.95	13.22/10.58	12.68/10.41
Proposed	**11.44/5.86**	**6.82/4.13**	**6.75/3.69**	**28.81/13.01**	**25.63/8.53**	**22.18/8.04**	**13.10/10.68**	**8.03/6.74**	**7.14/6.12**
Optimal	7.46/3.67	4.20/2.19	3.45/1.97	23.65/8.00	21.12/6.37	20.53/5.27	7.75/6.03	5.46/3.86	4.57/3.51

TABLE VI

THE BMP OF THE RESULTANT DISPARITY MAP ON MID 2006 [39], MID 2014 [40], AND KITTI 2015 [33] DATASET WITH MC-CNN. THE BMP OF THE REFINED DISPARITY MAP USING GROUND TRUTH CONFIDENCE IS MEASURED AS 'OPTIMAL'. THE BMP IS MEASURED WITH ONE/THREE PIXEL ERRORS. THE RESULT WITH THE LOWEST BMP IN EACH EXPERIMENT IS HIGHLIGHTED

Dataset	MID 2006 [39]			MID 2014 [40]			KITTI 2015 [33]		
Validations	+mod.	+GCPs.	+AGCPs.	+mod.	+GCPs.	+AGCPs.	+mod.	+GCPs.	+AGCPs.
Initial disparity	17.04/8.31	17.04/8.31	17.04/8.31	39.56/20.69	39.56/20.69	39.56/20.69	20.62/15.79	20.62/15.79	20.62/15.79
Haeusler et al. [23]	13.76/7.10	11.11/6.89	9.49/6.13	35.41/15.83	32.16/13.58	31.18/11.48	16.05/12.16	14.06/10.68	12.31/10.34
Spyropoulos et al. [20]	13.28/7.02	10.83/6.57	9.62/5.86	34.12/15.46	31.68/12.96	30.64/11.02	15.87/11.84	12.95/8.84	11.31/8.61
Park and Yoon [24]	13.14/6.86	10.63/6.03	9.33/5.72	33.68/14.92	30.71/11.69	28.31/10.71	15.04/11.44	12.18/8.36	11.28/8.12
Poggi et al. (O(1)) [25]	11.90/6.39	9.82/5.85	8.28/5.15	32.29/13.75	29.65/10.00	27.19/9.44	13.95/10.97	11.67/8.12	10.46/7.97
Kim et al. [27]	12.41/6.18	9.94/5.12	8.35/4.93	32.48/13.76	28.11/9.92	26.47/9.42	14.16/10.83	10.84/7.79	9.72/7.54
Poggi et al. [29]	11.60/6.02	9.73/4.63	8.23/4.13	29.81/13.34	27.13/9.87	25.71/9.23	14.04/10.08	9.65/7.41	8.83/7.11
Seki et al. [28]	11.35/5.78	9.64/4.71	7.80/4.02	28.63/12.99	26.57/9.34	24.56/8.88	14.37/9.86	9.32/6.96	9.02/6.63
Shaked et al. [32]	13.29/6.92	12.14/6.92	10.46/6.11	35.97/16.23	31.86/13.92	30.76/11.12	16.47/11.18	14.04/9.57	12.05/10.13
Proposed	**10.65/5.65**	**8.79/4.03**	**6.90/3.71**	**26.71/10.98**	**23.19/8.78**	**22.18/8.15**	**13.16/9.12**	**8.03/6.18**	**7.80/5.98**
Optimal	5.92/2.11	4.55/1.84	4.02/1.56	21.87/7.70	18.79/5.87	17.61/4.54	7.29/5.94	4.30/3.65	3.59/3.31

VI. CONCLUSION

In this study, we have presented a learning framework for estimating the stereo confidence through CNNs. We have shown that the optimal confidence features can be learned from the matching probability together with the disparity map. The matching probability volume is first generated

by refining and normalizing the matching cost volume, and then the confidence is estimated through deep networks with the matching probability and its corresponding disparity as inputs. Highly discriminative confidence features are learned by leveraging the multi-scale disparity features. Moreover, the proposed semi-supervised loss enables us to effectively learn the networks even with sparse ground truth disparity maps by using highly confident pixels of intermediate results for computing the image reconstruction loss. We validated the effectiveness of the proposed method by obtaining accurate and robust disparity maps on public datasets and challenging outdoor scenes through the depth refinement procedure using the estimated confidence map.

REFERENCES

[1] 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. Workshop*, Nov. 2011, pp. 467–474.

[2] 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.*, vol. 114, no. 11, pp. 1180–1202, 2010.

[3] G. Egnal and R. P. Wildes, "Detecting binocular half-occlusions: Empirical comparisons of five approaches," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1127–1133, Aug. 2002.

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

[5] J. Kim, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1033–1040.

[6] Y. Heo, K. Lee, and S. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 807–822, Apr. 2011.

[7] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[8] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

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

[10] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1431–1442, Aug. 2008.

[11] S. Kim, B. Ham, B. Kim, and K. Sohn, "Mahalanobis distance cross-correlation for illumination-invariant stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1844–1859, Nov. 2014.

[12] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5695–5703.

[13] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3017–3024.

[14] K.-J. Yoon and I. S. Kweon, "Distinctive similarity measure for stereo matching under point ambiguity," *Comput. Vis. Image Understand.*, vol. 112, no. 2, pp. 173–183, 2008.

[15] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, "Cross-based local multipoint filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 430–437.

[16] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, Oct. 2006.

[17] G. Egnal, M. Mintz, and R. P. Wildes, "A stereo confidence metric using single view imagery with comparison to five alternative approaches," *Image Vis. Comput.*, vol. 22, no. 12, pp. 943–957, 2004.

[18] P. Mordohai, "The self-aware matching measure for stereo," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 1841–1848.

[19] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3033–3040.

[20] 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.*, Jun. 2014, pp. 1621–1628.

[21] A. Fusiello, V. Roberto, and E. Trucco, "Efficient stereo with multiple windowing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 858–863.

[22] X. Hu and P. Mordohai, "Evaluation of stereo confidence indoors and outdoors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1466–1473.

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

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

[25] 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.*, Oct. 2016, pp. 509–518.

[26] H. Hirschmüller, P. R. Innocent, and J. Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 229–246, 2002.

[27] S. Kim, D. Min, S. Kim, and K. Sohn, "Feature augmentation for learning confidence measure in stereo matching," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 6019–6033, Dec. 2017.

[28] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. Brit. Mach. Vis. Conf.*, vol. 10, Sep. 2016, pp. 1–13.

[29] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. Brit. Mach. Vis. Conf.*, vol. 10, Sep. 2016, pp. 1–13.

[30] A. Spyropoulos and P. Mordohai, "Correctness prediction, accuracy improvement and generalization of stereo matching using supervised learning," *Int. J. Comput. Vis.*, vol. 118, no. 3, pp. 300–318, 2016.

[31] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Nov. 2012.

[32] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6901–6910.

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

[34] 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.*, Oct. 2012, pp. 611–625.

[35] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.

[36] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6602–6611.

[37] L. Xu, Q. Yan, and J. Jia, "A sparse control model for image and video editing," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 197.

[38] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5638–5653, Dec. 2014.

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

[40] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, Sep. 2014, pp. 31–42.

[41] 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, pp. 992–996.

[42] 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, pp. 4541–4550.

[43] F. Tosi, M. Poggi, A. Tonioni, L. Di Stefano, and S. Mattoccia, "Learning confidence measures in the wild," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2017, pp. 1–13.

[44] 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, pp. 5238–5247.

[45] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2017, pp. 66–75.
[46] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 154–169.
[47] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1614–1622.
[48] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.
[49] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2015, pp. 689–692.
[50] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, 2016.

Sunok Kim (S'14) received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2014, where she is currently pursuing the joint M.S. and Ph.D. degrees in electrical and electronic engineering. Her current research interests include 3D image processing and computer vision, in particular, stereo matching, depth super-resolution, and confidence estimation.

Dongbo Min (M'09–SM'15) received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was a Post-Doctoral Researcher with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. From 2015 to 2018, he was an Assistant Professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been an Assistant Professor with the Department of Computer Science and Engineering, Ewha Womans University, Seoul. His current research interests include computer vision, deep learning, video processing, and continuous/discrete optimization.

Seungryong Kim (M'12) received the B.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2012 and 2018, respectively. He is currently a Post-Doctoral Researcher in electrical and electronic engineering at Yonsei University. His current research interests include 2D/3D computer vision, computational photography, and machine learning, in particular, sparse/dense feature descriptor and continuous/discrete optimization.

Kwanghoon Sohn (M'92–SM'12) received the B.E. degree in electronic engineering from Yonsei University, Seoul, South Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Member of the Research Engineer with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, South Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School of Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently an Underwood Distinguished Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing and computer vision.