Adversarial Confidence Estimation Networks for Robust Stereo Matching

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

***Abstract*— Stereo matching aiming to perceive the 3-D geometry of a scene facilitates numerous computer vision tasks used in advanced driver assistance systems (ADAS). Although numerous methods have been proposed for this task by leveraging deep convolutional neural networks (CNNs), stereo matching still remains an unsolved problem due to its inherent matching ambiguities. To overcome these limitations, we present a method for jointly estimating disparity and confidence from stereo image pairs through deep networks. We accomplish this through a minmax optimization to learn the generative cost aggregation networks and discriminative confidence estimation networks in an adversarial manner. Concretely, the generative cost aggregation networks are trained to accurately generate disparities at both confident and unconfident pixels from an input matching cost that are indistinguishable by the discriminative confidence estimation networks, while the discriminative confidence estimation networks are trained to distinguish the confident and unconfident disparities. In addition, to fully exploit complementary information of matching cost, disparity, and color image in confidence estimation, we present a dynamic fusion module. Experimental results show that this model outperforms the state-of-the-art methods on various benchmarks including real driving scenes.**

***Index Terms*— Stereo confidence, confidence estimation, generative adversarial network, dynamic feature fusion.**

I. Introduction

STEREO matching for reconstructing the 3-D geometric configuration of a scene is a key enabler to realize various tasks used in advanced driver assistance systems (ADAS), including simultaneous localization and mapping (SLAM) and 3-D reconstruction [1]–[5]. For decades, numerous methods have been proposed for this task by leveraging handcrafted [1], [6] and/or machine learning based techniques [7], [8]. However, they frequently fail to produce an accurate depth map due to challenging elements, such as reflective surfaces, textureless regions, repeated pattern regions, occlusions [9]–[11], and photometric deformations incurred by illumination and camera specification variations [12], [13]. Especially, in real driving circumstances, these limitations frequently appear, which degrades stereo matching by existing methods.

To overcome these challenges, most approaches adopted a post-processing step in a manner that a set of unreliable disparities is first determined using confidence measures and then refined using neighboring reliable disparities [14]–[17]. Conventionally, hand-designed confidence measures [14]–[18] followed by shallow classifiers [19], [20] have been used to predict the confidence, but they have shown limited performance. Recent approaches have attempted to estimate the confidence by leveraging deep convolutional neural networks (CNNs) thanks to their high robustness [21]–[29], and have shown highly improved performance in comparison to handcrafted methods [14]–[18]. However, those aforementioned techniques focus on estimating the confidence of a pre-determined initial disparity only, ignoring the possibility of improving both the disparity and confidence estimation performance simultaneously.

Recently, some methods [8], [29] proposed to improve the quality of both disparity and confidence through deep networks. The underlying assumption is that an improved disparity helps to estimate the confidence more accurately. They simultaneously train two sub-networks that consist of the cost aggregation and confidence estimation networks. Although disparity and confidence estimation performance can be improved gradually during training, they do not have an explicit mechanism that uses the confidence estimation networks to boost the cost aggregation networks.

Meanwhile, generative adversarial networks (GANs) [30] have achieved impressive results in numerous computer vision tasks [31]–[35] that generates perceptually realistic solutions. Some stereo matching approaches [36], [37] also adopted the adversarial learning to estimate perceptually realistic disparity maps. However, there are no studies investigating how the discriminator can be used as the confidence estimator.

In this paper, we introduce novel deep networks and a learning framework for overcoming the aforementioned

Manuscript received June 4, 2019; revised December 5, 2019 and February 18, 2020; accepted May 8, 2020. Date of publication June 3, 2020; date of current version November 1, 2021. This work was supported by the Institute for Information and Communications Technology Promotion (IITP) Grant funded by the Korean Government (MSIP), Development of the High-Precision Natural 3-D View Generation Technology Using Smart-Car Multi Sensors and Deep Learning, under Grant 2016-0-00197. The work of Dongbo Min was supported in part by the Young Researcher Program through the National Research Foundation of Korea (NRF) under Grant NRF-2018R1C1B6004622. The Associate Editor for this article was N. Zheng. *(Corresponding author: Kwanghoon Sohn.)*

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

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

Seungryong Kim is with the École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, and also with the Department of Computer Science and Engineering, Korea University, Seoul 02481, South Korea (e-mail: seungryong_kim@korea.ac.kr).

Digital Object Identifier 10.1109/TITS.2020.2995996

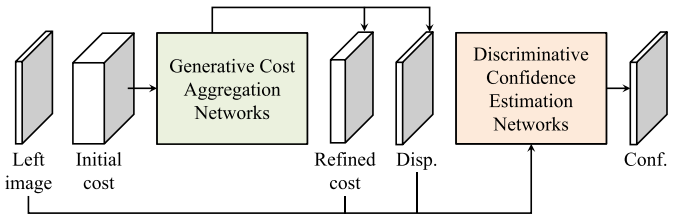


Fig. 1. Our overall network configuration that consists of two sub-networks, including generative cost aggregation networks and discriminative confidence estimation networks. Given a left color image and an initial matching cost, estimated by the existing matching cost computation methods [1], [7], as input, our networks output disparity and its confidence map.

limitations of existing deep CNN-based confidence estimation techniques. Inspired by GANs [30], the key idea is to design the cost aggregation and confidence estimation networks to be learned by the minmax optimization so that each network can be improved in an iterative and boosting fashion. Concretely, we formulate the two networks as two adversarial players, i.e., *generative cost aggregation* and *discriminative confidence estimation* networks as illustrated in Fig. 1. The generative cost aggregation networks are trained to accurately generate disparities at both confident and unconfident pixels from an input matching cost that are indistinguishable by the discriminative confidence estimation networks. At the same time, the discriminative confidence estimation networks are trained to distinguish the estimated confident and un-confident disparities. By training our networks in the minmax optimization fashion, both disparity and confidence maps, outputs of our networks, can be improved simultaneoulsy.

In addition, recent CNNs based stereo confidence estimation methods have been formulated by partially using single- or bi-modal inputs, e.g., matching cost only [8], disparity only [21], [22], matching cost and disparity [23], [29], or disparity and color [27], [28]. Moreover, a simple concatenation technique [38] is commonly used to fuse multi-modal confidence features, disregarding that the fusion weights may vary for each image depending on the attribute of confidence features. In order to fully exploit matching cost, disparity, and color image in confidence estimation, we present a dynamic fusion module in the discriminative confidence estimation networks in a way that an optimal fusion weight is dynamically determined conditioned on each input.

We verify the proposed confidence estimation method using aggregated ground control points (AGCPs) based propagation as in [29]. The proposed method is extensively evaluated through an ablation study and comparison with conventional handcrafted and deep CNNs-based methods on various benchmarks, including Middlebury 2006 [39], Middlebury 2014 [40], and KITTI 2015 [41].

We summarized the contributions of this paper as follows:

- We present an adversarial learning framework that jointly estimates disparity and confidence maps, where the discriminative confidence estimation network is formulated to operate as a discriminator to boost the generative cost aggregation network as well as generate the confidence.
- We present a novel confidence estimation network that fully exploits complementary information of matching cost, disparity, and color image through a dynamic fusion module.
- We conduct extensive experiments to evaluate our method on disparity and confidence estimation on various benchmarks.

The remainder of this paper is organized as follows. Sec. II describes the related works of our method. Sec. III analyses existing methods for disparity and confidence estimation and their limitations. Sec. IV presents the proposed network architectures and a learning framework. Sec. V introduces the validation method. Experimental results on disparity and confidence estimation are given in Sec. VI. Finally, conclusion and suggestions for future works are given in Sec. VII.

II. Related Works

A. Handcrafted Confidence Measures

Until the last few years, there have been extensive literatures in confidence estimation, mainly based on handcrafted confidence measures [11], [42], [43]. A comprehensive review by Hu and Mordohai [44] concluded that there is no single confidence feature that yields consistently optimal performance. To solve this limitation, there have been various approaches to benefit from the combination among a different set of confidence features, and train a shallow classifier such as random decision forest [14]–[16], [18]. Haeusler *et al.* [18] combined confidence features consisting of left-right consistency, image gradient, and disparity variance. A similar approach was also proposed in [14]. However, the performance of the aforementioned methods is still limited since the selected confidence features are not optimal. To select the set of optimal confidence features among multiple confidence features, Park and Yoon [15] proposed to utilize the machine learning technique that computes the importance of confidence features and train the regression forest classifier using selected confidence features. Poggi and Mattoccia [16] employed the set of confidence features from disparity map. While the aforementioned methods detect unconfident pixels in a pixel-level, Kim *et al.* [17] leveraged a spatial context to estimate the confidence in a superpixel-level. All of those methods used handcrafted confidence features, and thus they may not be optimal to detect unconfident pixels.

B. Deep CNN-Based Confidence Measures

With the recent progress of deep CNNs, several approaches have been proposed to measure the stereo confidence through deep CNNs [21]–[25], [27]–[29]. A quantitative evaluation of confidence measures that use shallow classifiers or CNN-based classifiers has been performed in [25]. Formally, these CNN-based methods first extract the confidence features from only disparity [21], [22], both disparity and cost volume [23], [29], and both disparity and color [28] to predict the confidence. In [21], a confidence estimation network was proposed that only takes a left disparity map as input. Seki and Pollefeys [22] proposed to use both left and right disparity maps in deep networks for improving the confidence prediction accuracy. In [24], the confidence map is refined

by leveraging a local consistency within an estimated confidence map. Tosi *et al.* [26] proposed a novel self-supervised strategy, which generates training labels by leveraging a pool of appropriately combined conventional confidence measures. There were attempts to benefit from a color image [27], [28]. They encode the local information from the color image to detect high frequency patterns, showing a performance gain in the confidence estimation. In spite of recent advances in the confidence estimation, the aforementioned methods estimate the confidence of an initial pre-determined disparity and do not consider improving its accuracy.

Recently, some methods [8], [29] attempted to improve the quality of disparity and its corresponding confidence simultaneously, under the assumption that an improved disparity helps to estimate confidence more accurately. They formulate two sub-networks consisting of cost aggregation networks and confidence estimation networks, and simultaneously learn them during training. Shaked and Wolf [8] jointly estimated both disparity and confidence maps. Kim *et al.* [29] also tried to estimate improved disparity and confidence maps through a unified deep network. However, they do not have an explicit mechanism to boost each network, where the confidence estimation networks cannot boost the cost aggregation networks. Contrarily, we formulate an adversarial learning framework to boost both the cost aggregation and confidence estimation networks.

C. Generative Adversarial Networks

GANs [30] have achieved impressive results in numerous computer vision and image processing applications, such as image colorization [31], [32], image super-resolution [33], image inpainting [34], and representation learning [35], enabling generating of perceptually realistic solutions. In disparity estimation literatures, some methods [36], [37], [45], [46] also exploited the adversarial learning to estimate perceptually realistic disparity maps. Reference [45] tried to solve a monocular depth estimation problem while our method tries to solve two-view stereo matching problem. Reference [46] tried to solve cross-spectral, i.e., color and infrared images, stereo matching, while our method considers stereo matching for two color images. Some methods [36], [37] tried to solve stereo mathing for two color images similar to ours, but the discriminator networks in these methods cannot operate as confidence estimator, and there are no studies to investigate how the discriminator networks can be used as the confidence estimator. Taking such a boosting mechnism into account is the topic of this paper.

III. Problem Statement

A. Confidence Estimation of Initial Disparity

Let us define a pair of stereo images $\{I^l, I^r\}$. The objective of stereo matching is to estimate a disparity D_i that is defined for each pixel $i = [i_{\mathrm{x}}, i_{\mathrm{y}}]^T$ between stereo image pairs. To this end, the matching costs $C_{i,d}$ between I_i^l and $I_{i'}^r$, where $i' = i - [d, 0]^T$, among disparity candidates $d = \{1, \ldots, d_{\max}\}$ are first measured, and then aggregated and/or optimized for determining the disparity D_i. Since most existing methods, including deep CNN-based methods [7], [47], cannot provide fully reliable solutions due to the inherent challenges of stereo matching, several approaches [14], [15], [17], [22], [23] employed an additional module to predict a confidence Q_i of the estimated disparity D_i.

Formally, the confidence estimation pipeline involves extracting confidence features from matching costs, disparity maps, and/or color images, and training the confidence predictor using the confidence features and ground-truth confidence maps. Unlike conventional approaches that use handcrafted features to train a shallow classifier [14]–[18], recent methods attempted to estimate the confidence by training deep CNNs such that the confidence features and predictors are trained simultaneously in an end-to-end manner [21]–[23]. The aforementioned approaches estimate only the confidence Q_i of a pre-determined initial disparity D_i, and adopt the subsequent disparity refinement scheme to improve the initial disparity D_i using the estimated confidence Q_i.

B. Existing Approaches for Unified Disparity and Confidence Estimation and Their Limitations

More recently, some methods [8], [29] have designed deep networks to simultaneoulsy improve the quality of disparity D_i and its confidence Q_i at each training iteration, and shown that an improved disparity helps to estimate its confidence more accurately. Concretely, these approaches formulate two modules as shown in Fig. 2(a), including cost aggregation networks that provide the refined cost C^t and disparity D^t at t-th iteration from an initial cost C, i.e., $\{C^t, D^t\} = G(C; W_G)$. For the simplicity of notation, we denote $C^t = G^C(C; W_G)$ and $D^t = G^D(C; W_G)$ for the refined cost and disparity, respectively,[1] and confidence estimation networks that estimate the confidence Q^t, i.e., $Q^t = F(C^t, D^t; W_F)$, where W_G and W_F represent network parameters, respectively. For the sake of simplicity, let us represent $G(\cdot; W_G)$ and $F(\cdot; W_F)$ as $G(\cdot)$ and $F(\cdot)$, respectively.

To learn these networks $\{G, F\}$ in an end-to-end manner, two loss functions can be used. First, the cost aggregation networks use a disparity regression loss with L_1 norm [29] with respect to a ground-truth disparity D^* such that

$$\mathcal{L}_{\mathrm{disp}}(G) = \mathbb{E}_{C\sim p_{data}(C)}[\|G^D(C) - D^*\|_1], \tag{1}$$

where it is defined for all C following the data distribution $p_{data}(C)$. Alternatively, $\mathcal{L}_{\mathrm{disp}}$ can also be defined as the cross-entropy loss [8] with C^t and D^*.

Second, the cross-entropy loss [8], [29] with respect to the ground-truth confidence Q^* is generally used to train confidence estimation networks. The confidence estimation network that distinguishes confident and unconfident samples is trained by maximizing the following energy function:

$$\begin{aligned}\mathcal{L}_{\mathrm{conf}}(G, F) = {} & \mathbb{E}_{C\sim p_{pos}(C)}[\log F(G(C))] \\ & + \mathbb{E}_{C\sim p_{neg}(C)}[\log(1 - F(G(C)))],\end{aligned} \tag{2}$$

[1]It should be noted that since D^t is derived from C^t using a soft-argmax function [8], [29], [47] with no trainable parameters, the network parameters of G^C and G^D are shared as W_G.

where the first and second terms are defined for positive (or confident) and negative (or unconfident) samples, i.e., $C \sim p_{pos}(C)$ and $C \sim p_{neg}(C)$, respectively. In addition, to build positive and negative samples, the ground-truth confidence is defined such that $Q^* = T(a; \rho)$ with a truncation function [29], where $T(a; \rho) = 1$ if $a < \rho$, 0 otherwise. Note that the ground-truth confidence Q^* is actively varying according to the estimated disparity D^t as evolving training iterations.

A total loss function is then defined as $\mathcal{L}_{\text{disp}}(G) + \lambda \mathcal{L}_{\text{conf}}(G, F)$ with λ controlling the relative importance of the two objectives, and the networks can be trained by minimizing each corresponding loss function:

$$\{G^*, F^*\} = \underset{G}{\text{argmin}}\, \mathcal{L}_{\text{disp}}(G) + \lambda \underset{G,F}{\text{argmax}}\, \mathcal{L}_{\text{conf}}(G, F). \quad (3)$$

Since two networks $\{G, F\}$ are trained in an iterative manner, an initial disparity D and confidence Q can be improved as D^t and Q^t. Although these methods [8], [29] have shown the state-of-the-art performance compared to existing confidence estimation methods [21], [22] that use the fixed ground-truth confidence Q^* (i.e., $D^t = D$) during training, they do not have an explicit mechanism such that confidence estimation networks F improve the cost aggregation networks G explicitly. To be specific, by maximizing $\mathcal{L}_{\text{conf}}(G, F)$ in (3), cost aggregation networks G generates positive and negative samples that are well *distinguishable* by confidence estimation networks F. However, there is no explicit mechanism that F cannot help to improve the ability of G to generate reliable matching cost and disparity maps.

IV. Proposed Method

A. Motivation and Overview

To alleviate the aforementioned limitations of existing methods [8], [29], we present novel network architectures and a learning technique. The proposed networks consist of two subnetworks, including *generative cost aggregation networks* G and *discriminative confidence estimation networks* F. Inspired by GANs [30], our objective is to formultate these two networks as two adversarial players so that each network can be improved in an iterative and boosting manner, as illustrated in Fig. 2(b). Concretely, the generative cost aggregation networks G are designed to train the mapping function from an initial cost C to refined costs C^t and disparity D^t at t-th iteration such that $\{C^t, D^t\} = G(C; W_G)$. At the same time, the discriminative confidence estimation networks F are trained to distinguish positive and negative samples of refined cost C^t, disparity map D^t, and color image I^l by estimating the confidence Q^t as the mapping: $Q^t = F(C^t, D^t, I^l; W_F)$.

Moreover, existing deep confidence estimation networks [8], [21]–[23], [29] have been formulated using matching cost only [8], disparity only [21], [22], unified matching cost and disparity [23], [29], or disparity and color image [27], [28] as input, and they cannot benefit from the joint use of tri-modal data, including matching cost, disparity, and color image. To fully exploit tri-modal data consisting of refined cost C^t, disparity D^t, and color image I^l, the discriminative confidence estimation networks have the dynamic fusion module where the optimal fusion weight is dynamically determined conditioned on each input.

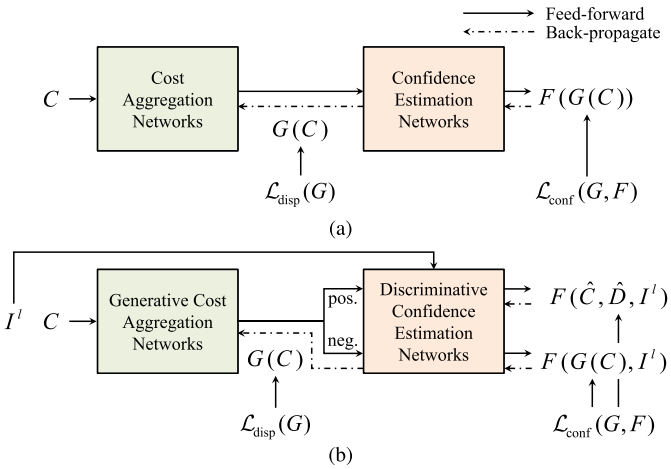


Fig. 2. Intuition of our networks: (a) conventional unified depth and confidence estimation methods [8], [29] and (b) ours. Unlike existing methods [8], [29], our networks, consisting of the generative cost aggregation and discriminative confidence estimation networks, are designed to be trained in the minmax optimization fashion. Discriminative confidence estimation networks are trained to distinguish generated positive samples (i.e., matching cost $\hat{C}^t$ and disparity $\hat{D}^t$) and negative samples using the loss function $\mathcal{L}_{\text{conf}}$. The derivative of $\mathcal{L}_{\text{conf}}$ with respect to negative samples is only back-propagated, which enables generative cost aggregation networks generate samples that are indistinguishable by the discriminative confidence estimation networks.

B. Network Architecture

1) Generative Cost Aggregation Networks: Similar to [29], generative cost aggregation networks consist of a residual convolutional module, a normalization layer, a top-K pooling layer, and a soft-argmax layer, as shown in Fig. 3. Specifically, the networks first aggregate the initial matching cost C using the encoder-decoder networks with skip layers similar to U-Net [48], consisting of sequential convolutional layers followed by batch normalization (BN) and rectified linear units (ReLU). We sequentially apply 2×2 max-pooling operators, resulting in a total down-sampling factor of 4. In the decoding parts, the intermediate features are upsampled using bilinear deconvolutional filters [48], [49], and concatenated with corresponding encoder features using skip layers. We compute a residual matching cost and estimate the aggregated matching cost R^t at t-th iteration. Note that the network parameters W_G are defined only on the residual cost aggregation module and there are no trainable parameters at subsequent layers. We then use the normalization layer that generates the matching probability volume P^t at t-th iteration to deal with the scale variation problems within matching cost volume [23], [47] as follows:

$$P^t_{i,d} = \frac{\exp(-R^t_{i,d}/\sigma)}{\sum_u \exp(-R^t_{i,u}/\sigma)}, \quad (4)$$

where $u = \{1, \ldots, d_{\max}\}$, and σ is a parameter for adjusting the flatness of the matching cost.

Furthermore, to deal with redundant parts in P^t that distract the performance of confidence estimation, we also use a top-K

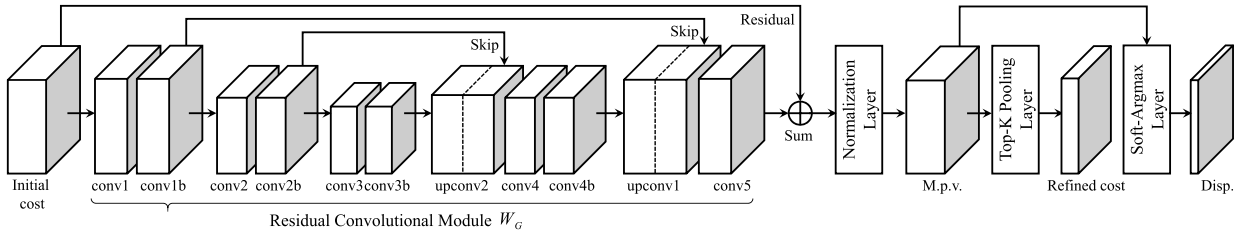


Fig. 3. Illustration of the generative cost aggregation networks G, consisting of a residual convolutional module, a normalization layer, a top-K pooling layer, and a soft-argmax layer.

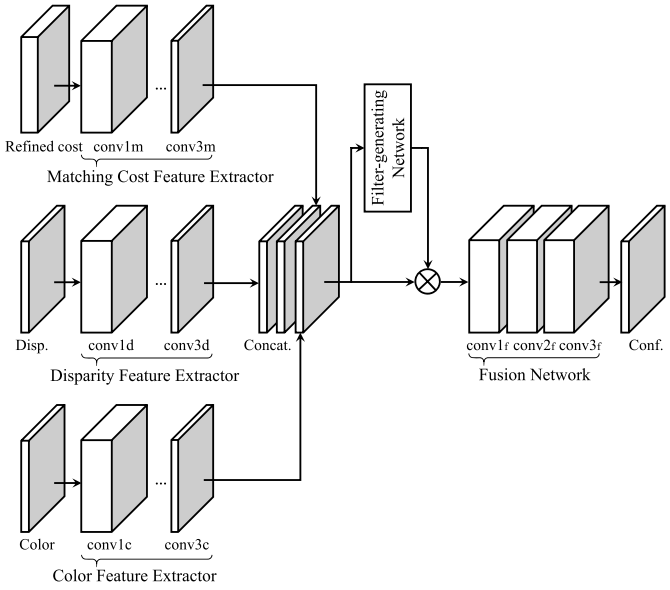


Fig. 4. Illustration of the discriminative confidence estimation networks F, consisting of tri-modal feature extractors for matching cost, disparity, and color image, an adaptive fusion module, and a confidence estimator.

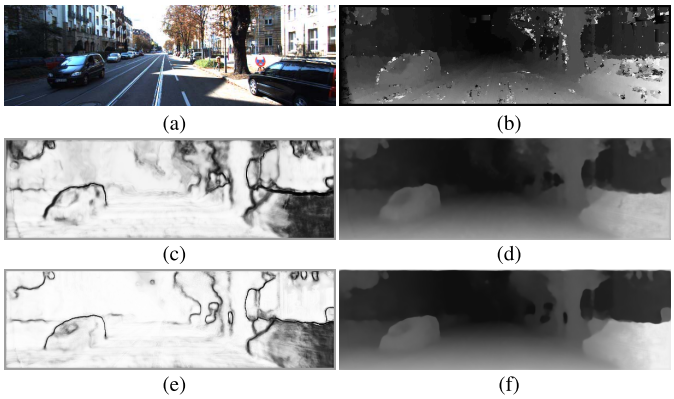


Fig. 5. The effect of the color feature extractor in our confidence estimation networks: (a) a left color image, (b) an initial disparity map estimated using MC-CNN [7], (c) an estimated confidence map using cost and disparity as inputs, (d) a refined disparity map by using (c) AGCPs-based propagation, (e) an estimated confidence map using cost, disparity, and color as inputs, and (f) a refined disparity map by using (e) AGCPs-based propagation. Using color feature extractor, the confidence estimation networks produce the result well aligned with the color image. (Best viewed on electronic version).

pooling layer [29], where P^t is projected into a fixed length. The refined matching probability C^t is obtained as follows:

$$C^t_{i,k} = \max^k_d P^t_{i,d}, \tag{5}$$

where $\max^k(\cdot)$ is the k-th maximal value for $k = \{1, \ldots, K\}$.

Finally, to estimate the current disparity D^t at t-th iteration, we use the soft-argmax layer [29], [47] as follows:

$$D^t_i = \sum\nolimits_d P^t_{i,d} d. \tag{6}$$

2) Discriminative Confidence Estimation Networks: Discriminative confidence estimation networks take tri-modal data consisting of matching cost[2] C^t, disparity D^t, and color image I^l as input, and output confidence Q^t, as shown in Fig. 4. Especially, unlike [29], our confidence estimation networks additionally use color information, enabling us to leverage a spatial context and estimate edge-preserved confidence maps [17]. Concretely, the discriminative confidence estimation networks first extract convolutional features f_{C^t}, f_{D^t}, and f_{I^l} from C^t, D^t, and I^l using network parameters W^C_F, W^D_F, and W^I_F, respectively, and then fuse them. Fig. 5 shows the effectiveness of the proposed confidence estimation. Similar to the generative cost aggregation networks, the discriminative confidence estimation networks also consist of sequential convolutional filters, followed by BN and ReLU. The pooling operation is not used to preserve the spatial resolution in this network.

In literatures [29], [50], a simple concatenation approach has been commonly used to fuse multi-modal features. However, such a simple approach often fails to generate optimal confidence features, since the fusion weights are fixed without considering test data at inference. Namely, the optimal fusion weights of the tri-modal features may vary depending on their attribute, but the simple concatenation technique is unable to consider such a dynamic fusion. To alleviate this limitation, inspired by [51], we introduce a dynamic fusion module for a dynamic combination of f_{C^t}, f_{D^t}, and f_{I^l}, where the optimal fusion weight $W^{H,*}_F$ is trained with respect to each input with an additional convolutional network, called filter-generating network, such that $W^{H,*}_F = H(f_{C^t}, f_{D^t}, f_{I^l}; W^H_F)$ with network parameters W^H_F. Unlike the fixed fusion weights as in [29], $W^{H,*}_F$ is estimated conditioned on input, thus enabling more optimal fusion. The confidence is finally estimated through fusion networks with parameters W^P_F. The confidence

[2]Note that the input of the confidence estimation network is the matching probability, but we refer to it as matching cost for clarity.

Q^t can be estimated as

$$\begin{aligned} Q^t &= F(C^t, D^t, I^l; W_F) \\ &= F(f_{C^t}, f_{D^t}, f_{I^l}; H(f_{C^t}, f_{D^t}, f_{I^l}|W_F^H), W_F^P), \end{aligned} \quad (7)$$

where $W_F = \{W_F^C, W_F^D, W_F^I, W_F^H, W_F^P\}$.

C. Loss Functions

1) Loss for Generative Cost Aggregation Networks: A major challenge to train stereo matching networks is the lack of dense ground-truth disparity maps especially for outdoor scenes such as real driving scenes. The synthetic outdoor data with ground-truth disparity maps is publicly available, but they incur the domain adaption issue [52]. Some benchmarks [53] provide outdoor data taken with depth sensors, e.g., LiDAR, but sparse disparity maps are not suitable for training the networks in a fully supervised manner. To overcome this issue, we use an unsupervised loss function using an image reconstruction loss [54] as well as a supervised loss function based on L_1 norm to directly regress the disparity map, similar to [29]. The loss function $\mathcal{L}_{\text{disp}}(G)$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{disp}}(G) = \mathbb{E}_{C\sim p_{data}(C)}[\|G^D(C) - D^*\|_1] \\ + \mathbb{E}_{C'\sim p_{data}(C')}[\|I^r(G^D(C')) - I^l\|_1], \end{aligned} \quad (8)$$

where $I^r(G^D(C'))$ is a warped right color image using the estimated disparity $G^D(C')$, implemented by a bilinear sampler [49] that enables end-to-end learning. While C is only defined for sparse ground-truth disparity D^*, C' is defined for all pixels. It should be noted in our method, by learning generative cost aggregation networks G with the loss function $\mathcal{L}_{\text{conf}}$ for confidence estimation networks, which will be described in the following, as well as $\mathcal{L}_{\text{disp}}$, more perceptually plausible disparity maps can be estimated.

2) Loss for Discriminative Confidence Estimation Networks: To learn the discriminative confidence estimation networks, we employ the cross-entropy loss function [8], [29] with respect to the ground-truth confidence map. However, unlike existing methods [8], [29], we design the loss function $\mathcal{L}_{\text{conf}}(G, F)$ to adversarially train two networks G and F. Namely, the discriminative confidence estimation networks F are trained to distinguish positive and negative samples (i.e., matching cost and disparity) more discriminatively, while the generative confidence estimation networks G are trained to generate negative samples that have the distribution more similar to that of positive samples as evolving training iterations [30]. Following this design strategy, the loss function $\mathcal{L}_{\text{conf}}(G, F)$ is defined as follows:

$$\begin{aligned} &\mathcal{L}_{\text{conf}}(G, F) \\ &= \mathbb{E}_{\hat{C}^t\sim p_{pos}(C^t), \hat{D}^t\sim p_{pos}(D^t), I^l\sim p_{pos}(I^l)}[\log F(\hat{C}^t, \hat{D}^t, I^l)] \\ &\quad + \mathbb{E}_{C\sim p_{neg}(C), I^l\sim p_{neg}(I^l)}[\log(1 - F(G(C), I^l))], \end{aligned} \quad (9)$$

where the first and second terms are defined for positive and negative samples, i.e., $C \sim p_{pos}(C)$ and $C \sim p_{neg}(C)$, respectively. $\hat{C}^t$ and $\hat{D}^t$ represent generated positive samples from generative cost aggregation networks G determined with respect to the ground-truth confidence Q^*. Although positive samples $\hat{C}^t$ and $\hat{D}^t$ are estimated from G, the derivative of $\mathcal{L}_{\text{conf}}$ with respect to G for positive samples cannot be backpropagated, while the derivative for negative samples is only backpropagated. As exemplified in Fig. 6, by using the confidence loss $\mathcal{L}_{\text{conf}}$, our method provides more plausible intermediate disparity and confidence maps.

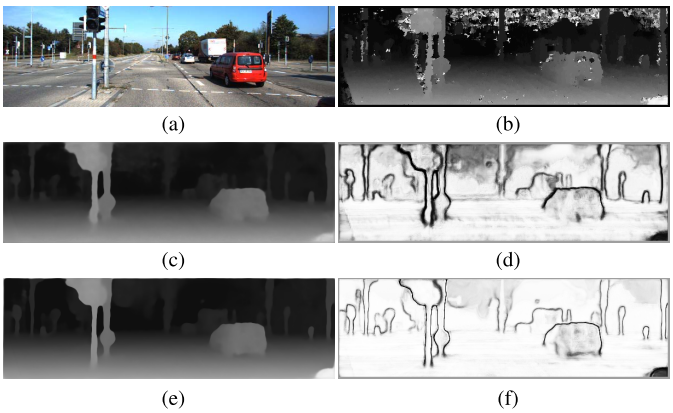


Fig. 6. The effects of the proposed adversarial learning technique: (a) a left color image, (b) an initial disparity map estimated using MC-CNN [7]. (c) and (d) are an intermediate disparity map and an estimated confidence map without adversarial confidence learning, (e) and (f) are obtained with adversarial confidence learning. By learning the disparity and confidence maps in an adversarial manner, our method provides more plausible intermediate disparity and confidence maps. (Best viewed on electronic version).

3) Full Objectives: Similar to GANs [30], we formulate the minmax optimization to learn our full networks consisting of generative cost aggregation networks and discriminative confidence estimation networks. The total loss fuction $\mathcal{L}_{\text{total}}$ is defined as follows:

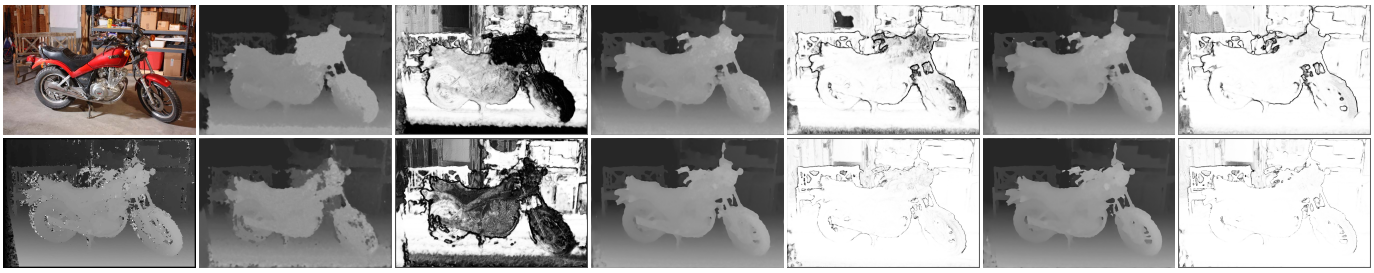
$$\mathcal{L}_{\text{total}}(G, F) = \mathcal{L}_{\text{disp}}(G) + \lambda\mathcal{L}_{\text{conf}}(G, F), \quad (10)$$

and the networks are then optimized in an adversarial fashion such that

$$\begin{aligned} \{G^*, F^*\} &= \operatorname*{argmin}_G \max_F \mathcal{L}_{\text{total}}(G, F) \\ &= \operatorname*{argmin}_G \mathcal{L}_{\text{disp}}(G) \\ &\quad + \lambda \operatorname*{argmin}_G \max_F \mathcal{L}_{\text{conf}}(G, F). \end{aligned} \quad (11)$$

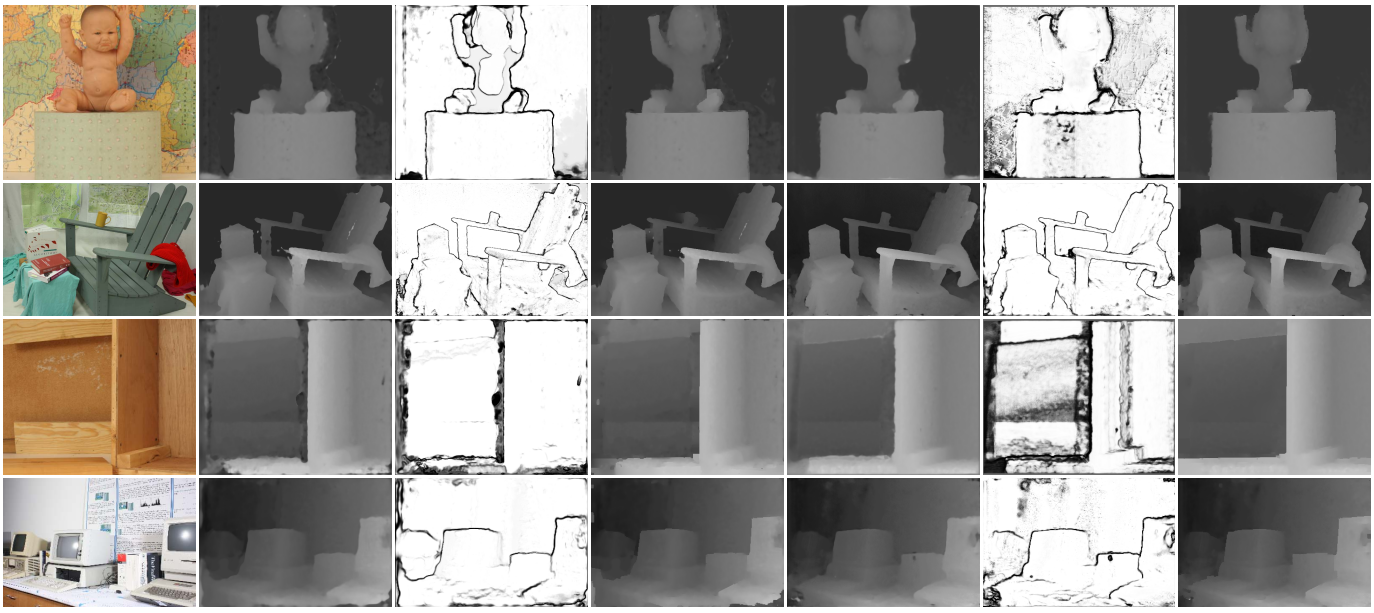
Within this optimization, G is trained to generate the disparity not only more similar to the ground-truth disparity but also indistinguishable by F, thus providing more reliable and error-reduced disparity. With progressively improved hard confident and unconfident samples, the ability of confidence estimation of F can also be improved. Note that unlike the previous methods [8], [29] where the confidence estimation networks cannot boost the disparity estimation performance, our method mutually boosts the diparity and confidence estimation.

In practice, at an early stage of training, generated negative samples are clearly different from generated positive samples, and thus discriminative confidence estimation networks can easily distinguish the generated samples. In this case, $\log(1 - F(G(C), I^l))$ in $\mathcal{L}_{\text{conf}}$ easily saturates. To solve this, rather than training G by minimizing $\log(1 - F(G(C), I^l))$, we train G to maximize $\log F(G(C), I^l)$ similar to [30].



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

Fig. 7. Visualizations of intermediate disparity and confidence maps as evolving training iterations on MID 2014 dataset [40]: (a) a color image and an initial disparity, and results by (top) Kim *et al.* [29] and (bottom) Ours at each epoch of 50 in (b), (c), 100 in (d), (e), and 150 in (f), (g).



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

Fig. 8. Comparisons of unified depth and confidence estimation methods on MID 2006 [39] and MID 2014 [40] datasets, using census-SGM (first two raws) and MC-CNN (last two rows): (a) color images, and final disparity, confidence, and refined disparity maps with propagation of (b), (c), (d) Kim *et al.* [29] and (e), (f), (g) Ours. Our method has shown the robustness againt challenging regions such as reflective surfaces and textureless regions.

The training optimization is then reformulated as follows:

$$\{G^*, F^*\} = \underset{G}{\operatorname{argmin}}\, \mathcal{L}_{\text{disp}}(G) + \lambda \underset{F}{\operatorname{argmax}}\, \mathcal{L}_{\text{conf}}(G, F) + \lambda \underset{G}{\operatorname{argmax}}\, \mathbb{E}_{C\sim p_{neg}(C), I^l\sim p_{neg}(I^l)}[\log F(G(C), I^l)], \tag{12}$$

This objective function results in the same fixed point of the dynamics of G and F but provides much stronger gradients at early stages of training.

V. Validation

In this section, we introduce the validatation method to verify the effectiveness of our confidence estimation method in the post-processing step of the stereo matching pipeline. Concretely, as described in [29], the predicted confidence can be incorporated in AGCPs-based propagation as in [55]. First, we set GCPs that are classified as confident pixels. Then, we globally propagate the GCPs through an Markov random field (MRF)-based optimization. We utilize an aggregated data term to mitigate propagation errors by inaccurately estimated confidences. It was shown in [55] that a more robust data constraint using an aggregated data term leads to a better quality in the sparse data interpolation. In this context, we define the energy function for refined disparity map $\bar{D}$ according to final disparity map D' (i.e., D^t after convergence) as follows:

$$\sum_i \left(\sum_{v\in\mathcal{M}_i} h_v c^I_{i,v} (\bar{D}_i - D'_v)^2 + \gamma \sum_{j\in\mathcal{N}_i^4} w^I_{i,j} (\bar{D}_i - \bar{D}_j)^2 \right), \tag{13}$$

where $\mathcal{M}_i$ represents a set of neighborhoods, and is not limited to a 4-neighborhood, but more neighbors are used for ensuring a large support. We define $c^I_{i,v}$ using a bilateral kernel between pixel i and v in the feature space consisting of color intensity I and spatial location. h_v is the binary mask to indicate the GCPs. Similar to $c^I_{i,v}$, $w^I_{i,j}$ is the affinity between i and j in the feature space consisting of color intensity I and spatial location, and $\mathcal{N}_i^4$ represents a 4-neighborhood. γ controls the

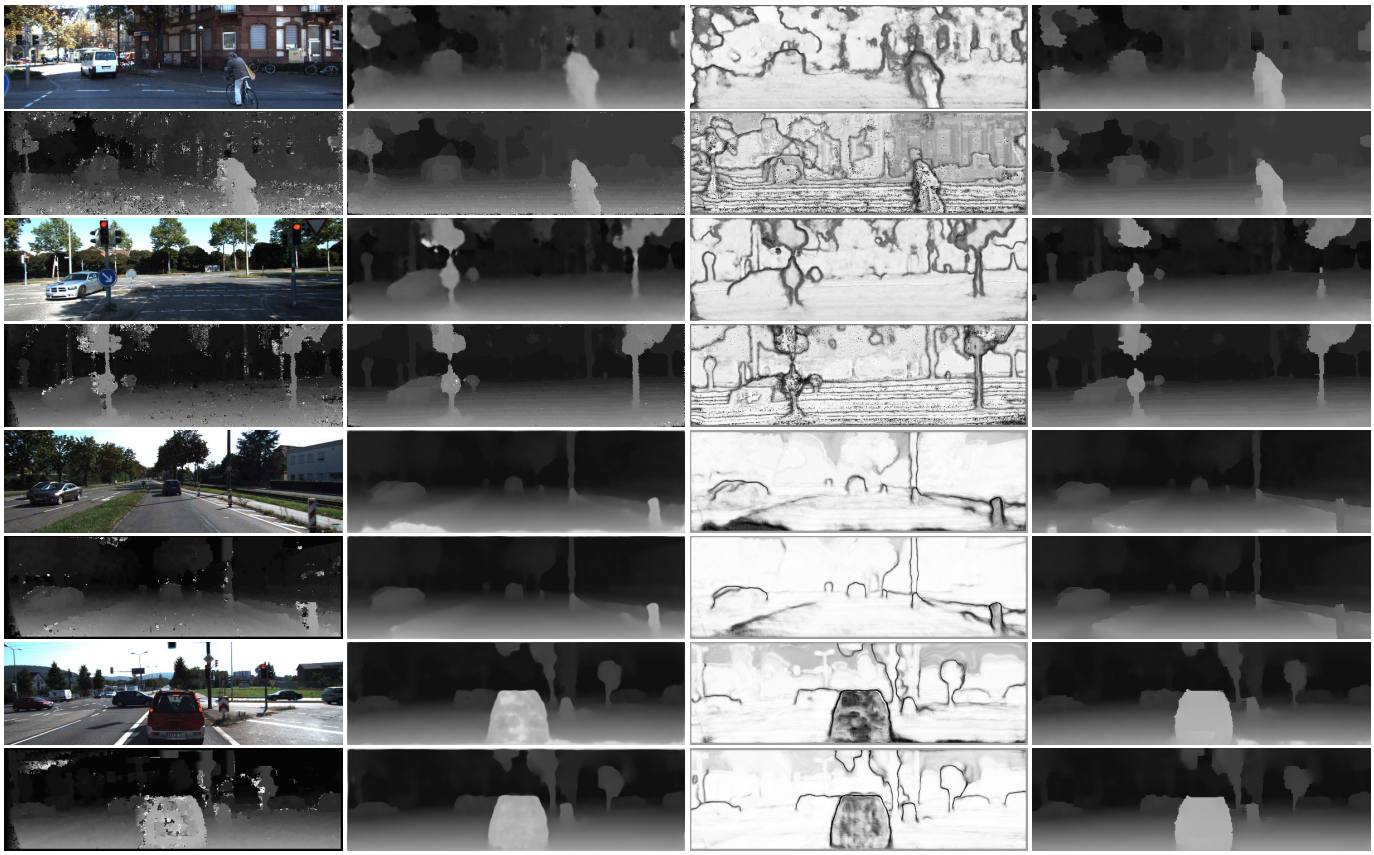


Fig. 9. Comparisons of unified depth and confidence estimation methods on KITTI 2015 [41], using census-SGM (1-4 rows) and MC-CNN (5-8 rows): (a) color images and disparity, (b) final disparity, (c) confidence, and (d) refined disparity maps with propagation of (top) Kim *et al.* [29] and (bottom) Ours. Our method has shown consistently reliable performances on pole, bicycle, and pedestrian regions.

relative importance of data and smoothness terms. This simple quadratic optimization can be efficiently solved using [55].

VI. Experimental Results

A. Training Details

The proposed method was implemented in MATLAB with `VLFeat MatConvNet toolbox` [56] and simulated on a PC with TitanX GPU. We make use of the stochastic gradient descent with momentum, and set the learning rate to 1×10^{-5} and the batch size to 20. To compute an initial matching cost, we used a census transform with a 5×5 local window and MC-CNN [7], respectively. For the census transform, we applied SGM [1] on estimated cost volumes by setting $P_1 = 0.008$ and $P_2 = 0.126$ as in [15]. For computing the MC-CNN, 'KITTI 2012 fast network' was used, provided at the author's website [57]. We set the threshold ρ as 0.9, and σ as 100 and 0.05 for census-SGM and MC-CNN, respectively.

B. Experimental Setup

In the following, we evaluated the proposed method in comparison to conventional shallow classifier-based approaches, such as Haeusler *et al.* [18], Spyropoulos *et al.* [14], Park and Yoon [15], Poggi and Mattoccia [16], Kim *et al.* [17], and CNNs-based approaches using disparity only, such as Poggi and Mattoccia (CCNN) [21], Seki and Pollefey (PBCP) [22], matching cost only, such as Shaked and Wolf [8], both disparity and matching cost, such as Kim *et al.* [29], and both color and disparity, such as Fu *et al.* (LFN) [28]. We obtained the results of [15], [17], and [29] by using the author-provided code, while the results of [8], [14], [18], [22], [28] were obtained by our own implementation. We re-implemented methods of [16] and [21] based on the author-provided code.

Following experimental settings of [29], we seperately trained our networks for indoor and outdoor datasets. For an indoor setting, we used 80 stereo pairs on the MPI dataset [58] for training, and evaluated the trained networks on 21 stereo pairs on the Middlebury 2006 (MID 2006) dataset [39] and 13 stereo pairs on the Middlebury 2014 (MID 2014) dataset [40]. We cropped the training images into 256×256-sized patches and totally obtained about 100,000 patches. For an outdoor setting, we used 194 stereo pairs on KITTI 2012 dataset [41] for the sparse supervised loss and 40,000 stereo pairs for the dense unsupervised loss in $\mathcal{L}_{\text{disp}}$ during training, and evaluated the trained networks on 200 stereo pairs on KITTI 2015 dataset [41] and Cityscapes dataset [59]. Note that to optimize the loss function $\mathcal{L}_{\text{disp}}$, consisting of supervised- and unsupervised losses, we used both supervised- and unsupervised losses after performing 200 epoches with the supervised loss only.

TABLE I

THE BMP OF THE FINAL DISPARITY MAP ON MID 2006 [39], MID 2014 [40], AND KITTI 2015 [41] DATASET WITH CENSUS-BASED SGM. THE BMP IS MEASURED WITH ONE AND THREE PIXEL ERRORS. THE RESULT WITH THE LOWEST BMP IN EACH EXPERIMENT IS HIGHLIGHTED

Datasets	MID 2006 [40]				MID 2014 [41]				KITTI 2015 [42]			
	Census-SGM		MC-CNN		Census-SGM		MC-CNN		Census-SGM		MC-CNN	
	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px
Shaked et al. [8]	10.16	8.12	13.58	8.66	25.79	14.23	28.56	15.02	13.32	11.45	14.08	7.94
Kim et al. [29]	8.32	5.86	9.12	5.49	24.65	11.08	25.69	12.01	9.27	8.67	10.51	8.24
Ours w/Sup.	7.92	5.20	8.57	5.03	22.84	9.76	23.97	11.90	8.81	8.12	9.72	8.03
Ours w/Unsup.	8.13	5.43	8.95	5.14	23.06	10.22	24.54	12.05	9.11	8.58	10.23	8.17
Ours	**7.88**	**5.05**	**8.49**	**4.93**	**22.23**	**9.45**	**23.27**	**11.69**	**8.41**	**7.85**	**9.12**	**7.76**
Shaked et al. [8] w/AGCPs	9.15	5.15	11.03	6.10	23.92	12.87	25.24	12.68	11.71	9.58	12.15	6.41
Kim et al. [29] w/AGCPs	6.75	3.69	6.90	3.71	22.18	8.04	22.18	8.15	7.14	6.12	7.80	5.98
Ours w/AGCPs	**5.84**	**2.66**	**4.87**	**2.51**	**20.59**	**7.17**	**20.36**	**7.39**	**6.97**	**2.66**	**5.46**	**4.05**

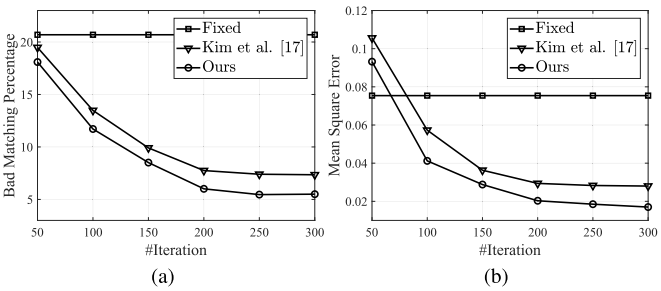


Fig. 10. The effectiveness of unified depth and confidence estimation of our method in comparison to Kim *et al.* [29]: (a) BMP of intermediate disparity maps and (b) MSE of estimated confidence maps as evolving training iterations. Note that 'Fixed' estimates only confidence without refining matching cost in our method.

To evaluate the performance of confidence estimation quantitatively, we used the sparsification curve and its area under curve (AUC) as used in [14], [15], [18], [22]. 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. For the higher accuracy of the confidence measure, its AUC value is lower. In addition, we measured mean squared errors (MSE) between estimated confidence and ground-truth confidence. To evaluate the disparity refinement performance, we also measured the bad matching percentage (BMP) as in [39]. Note that the BMP was obtained by measuring the ratio of erroneous pixels.

C. Unified Depth and Confidence Estimation Analysis

In order to demonstrate the synergistic effects of the proposed framework that simultaneously generates disparity and confidence maps, we first analyzed the convergence in comparison to the existing unified method [29]. Fig. 7 shows the qualitative results by evolving iterations. Fig. 8 and Fig. 9 show qualitative results of our method in comparison to [29]. For quantitative evaluations, we measured the average BMP of intermediate disparity maps at every 50 epochs, as shown in Table I and Fig. 10(a). Note that the conventional methods [21]–[23] estimate confidence map on fixed disparity map, while the proposed method as well as [8], [29] generates the intermediate disparity map by refining initial cost volumes and predicts the confidence map accordingly. In [8], [29], the BMP decreased as evolving the number of iterations, but was saturated. The proposed method shows the lowest BMP, demonstrating the superiority of the proposed method. In addition, by jointly using supervised and unsupervised disparity loss functions, our method has shown highly improved performance. We further analyzed the confidence estimation performance as evolving the iterations. For quantitative evaluations, we measured the MSE between estimated confidence and ground-truth confidence. Fig. 10(b) shows the MSE on MID 2014 dataset [40] obtained by MC-CNN for evolving epochs,[3] which demonstrates the superior performance of our confidence estimation. Especially, our method has shown the robustness against challenging regions such as reflective surfaces and textureless regions, as shown in Fig. 8, and consistently reliable performances on pole, bicycle, and pedestrian regions, as shown in Fig. 9.

D. Confidence Estimation Analysis

1) Ablation Study: We then analyzed our confidence estimation networks with ablation evaluations, with respect to color feature extractor and dynamic fusion module. For quantitative evaluations, we measured the average AUC values for various set of inputs and fusion methods.

First of all, ablation experiments to validate the effects of color feature extractor show that the confidence estimator can be improved by extracting the color feature as shown in Table II. Qualitative results also show the effectiveness of color feature extractor. Fig. 5(d) and Fig. 5(f) exemplify the refined disparity with the estimated confidences that use bi-modalities (cost volume and disparity) and tri-modalities (cost volume, disparity, and color), respectively. Secondly, we evaluated two different fusion methods; simple concatenation and dynamic fusion. Table II shows the effects of the adaptive weight learned with dynamic fusion module.

2) Comparison to Other Methods: 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 [14]–[18] and CNN-based methods [8], [21], [22], [28]. In these experiments, we only

[3]In these experiments, we did not measure the sparsification curve and AUC since the intial disparities are actively varied as evovling iterations.

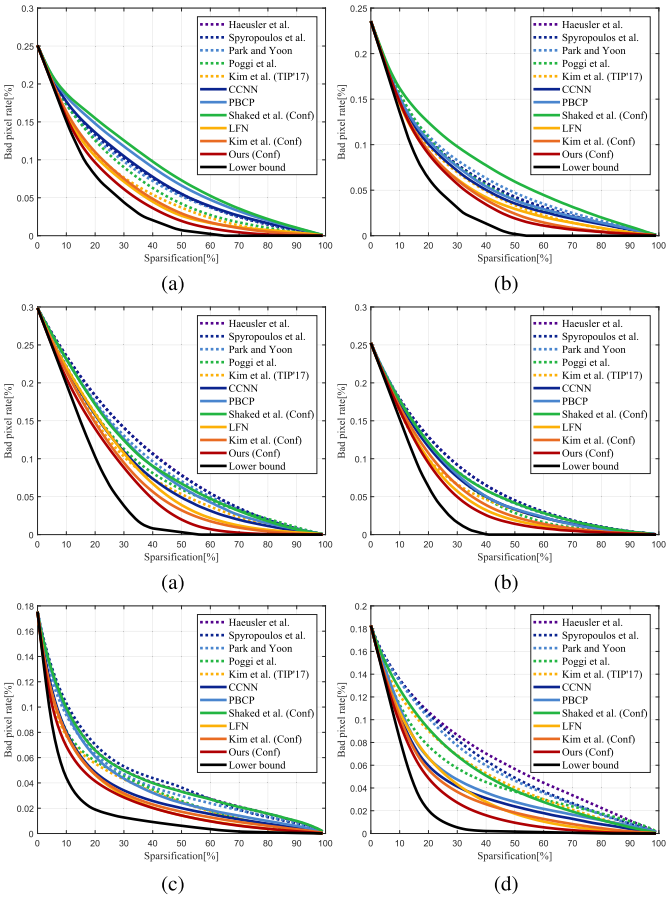


Fig. 11. Comparisons of sparsification curves for MID 2006 [39] using (a) census-SGM and (b) MC-CNN, MID 2014 [40] using (c) census-SGM and (d) MC-CNN, and KITTI 2015 dataset [41] using (e) census-SGM and (f) MC-CNN, respectively. The sparsification curve for the ground-truth confidence map is described as 'Lower bound'.

TABLE II

ABLATION STUDY FOR EACH COMPONENT OR OUR METHOD ON MID 2006 [39], MID 2014 [40], AND KITTI 2015 [53] DATASET, WHEN THE INITIAL MATCHING COST IS OBTAINED USING (TOP) CENSUS-SGM [1] AND (BOTTOM) MC-CNN [7]

Methods		[29]	[28]	Ours wo/col.	Ours wo/dyn.	Ours
Matching cost		✓		✓	✓	✓
Disparity		✓	✓	✓	✓	✓
Color			✓		✓	✓
Concatenation		✓	✓		✓	
Dynamic fusion				✓		✓
Census -SGM	MID 2006	0.0417	0.0416	0.0416	0.0413	**0.0411**
	MID 2014	0.0730	0.0752	0.0725	0.0712	**0.0702**
	KITTI 2015	0.0403	0.0405	0.0398	0.0399	**0.0391**
MC -CNN	MID 2006	0.0389	0.0393	0.0385	0.0382	**0.0376**
	MID 2014	0.0692	0.0692	0.0690	0.0690	**0.0688**
	KITTI 2015	0.0247	0.0253	0.0242	0.0240	**0.0231**

evaluated the confidence estimation module in our method, thus we denote our methods as 'Ours (Conf)'. Furthermore, for fair comparison, we also evaluated the confidence estimation performance only for [8], [29], i.e., Shaked *et al.* (Conf) [8] and Kim *et al.* (Conf) [29]. The lower bound of AUC can be obtained with a ground-truth confidence map. Sparsification

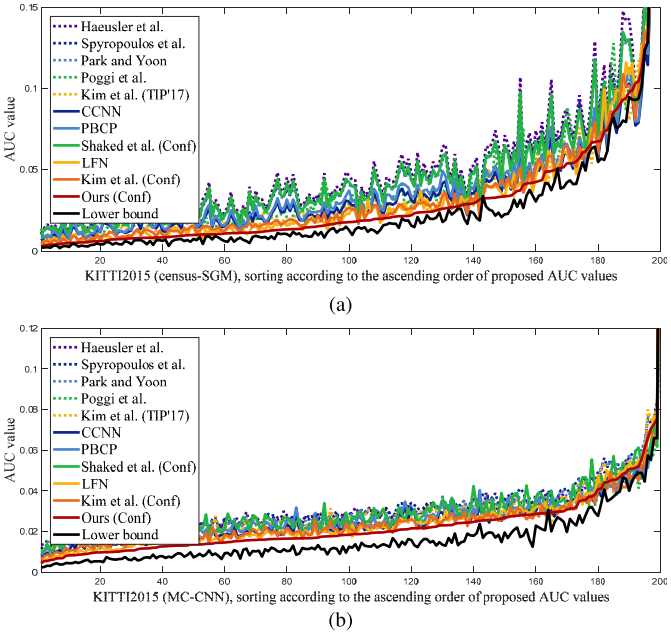


Fig. 12. Comparisons of AUC values for (a) census-based SGM and (b) MC-CNN for the KITTI 2015 dataset [41]. We sort the AUC values in the ascending order according to the AUC values. The 'Lower bound' of AUC values are computed using the ground-truth confidence map.

curves for MID 2014 dataset [40] and KITTI 2015 dataset [41] with census-based SGM and MC-CNN are shown in Fig. 11. The results have shown that the proposed confidence estimator exhibits a better performance than both conventional handcrafted approaches and CNN-based approaches. Fig. 12 describes the AUC values, which are sorted in ascending order, for the KITTI 2015 dataset [41] with census-based SGM and MC-CNN, respectively. The average AUC and MSE values with census-based SGM and MC-CNN for MID 2006, MID 2014, and KITTI 2015 datasets were summarized in Table III. The handcrafted approaches showed inferior performance than the proposed method due to low discriminative power of the handcrafted confidence features. CNN-based methods [8], [21], [22], [28] have improved confidence estimation performance compared to existing handcrafted methods such as [15], 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. Especially, the proposed method always yields the lowest AUC values, showing the superiority of the proposed method compared to the existing CNN-based methods [8], [21], [22], [28].

In addition, the estimated confidence maps are qualitatively shown in Fig. 13, Fig. 14, Fig. 15, and Fig. 16. Experimental results demonstrate that the proposed networks enable us to estimate more accurate disparity and confidence maps simultaneously in a boosting manner.

E. Depth Refinement Analysis

To verify the robustness of the confidence measures, we refined the disparity map using the confidence map estimated by several confidence measure approaches including ours. For refining the disparity maps, we used AGCPs-based

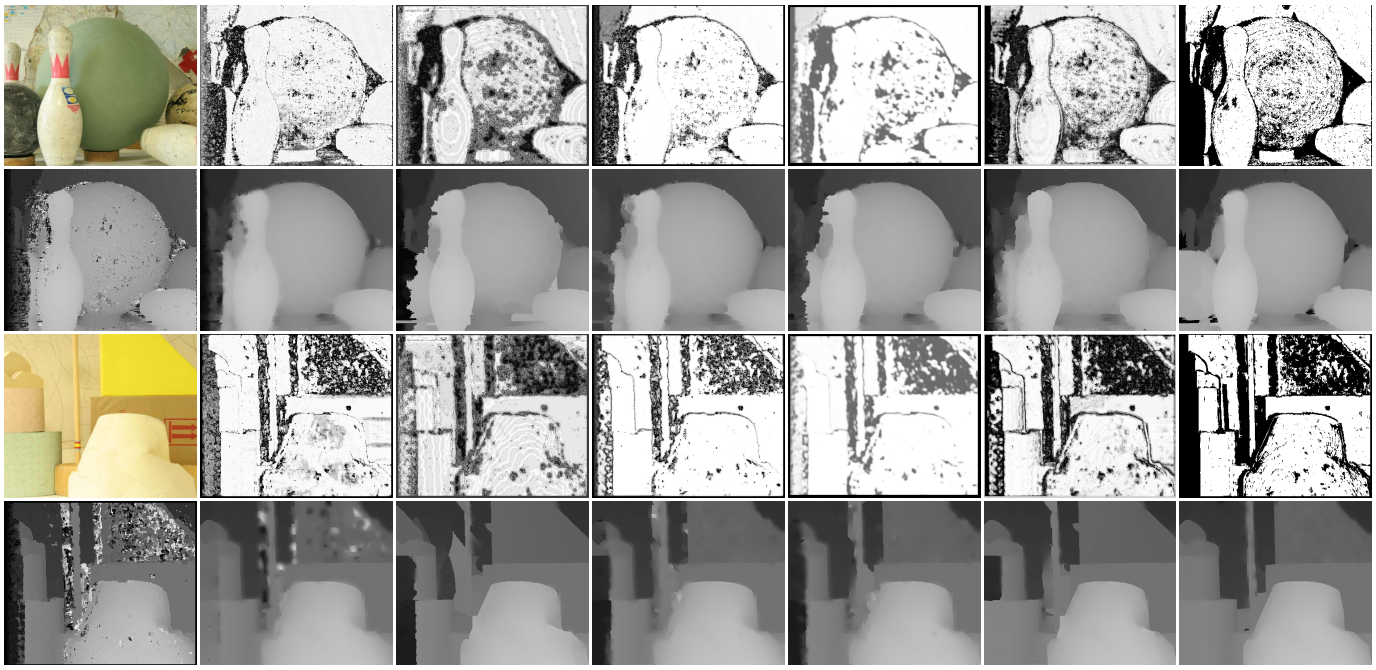


Fig. 13. The confidence and refined disparity maps on MID 2006 dataset [39] using census-SGM (first two rows), and MC-CNN (last two raws). (a) color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Park and Yoon [15], (c) Poggi and Mattoccia [16], (d) CCNN [21], (e) PBCP [22], (e) Ours (Conf), and (g) ground-truth confidence map. Our method has shown the robustness againt challenging regions such as reflective surfaces and textureless regions.

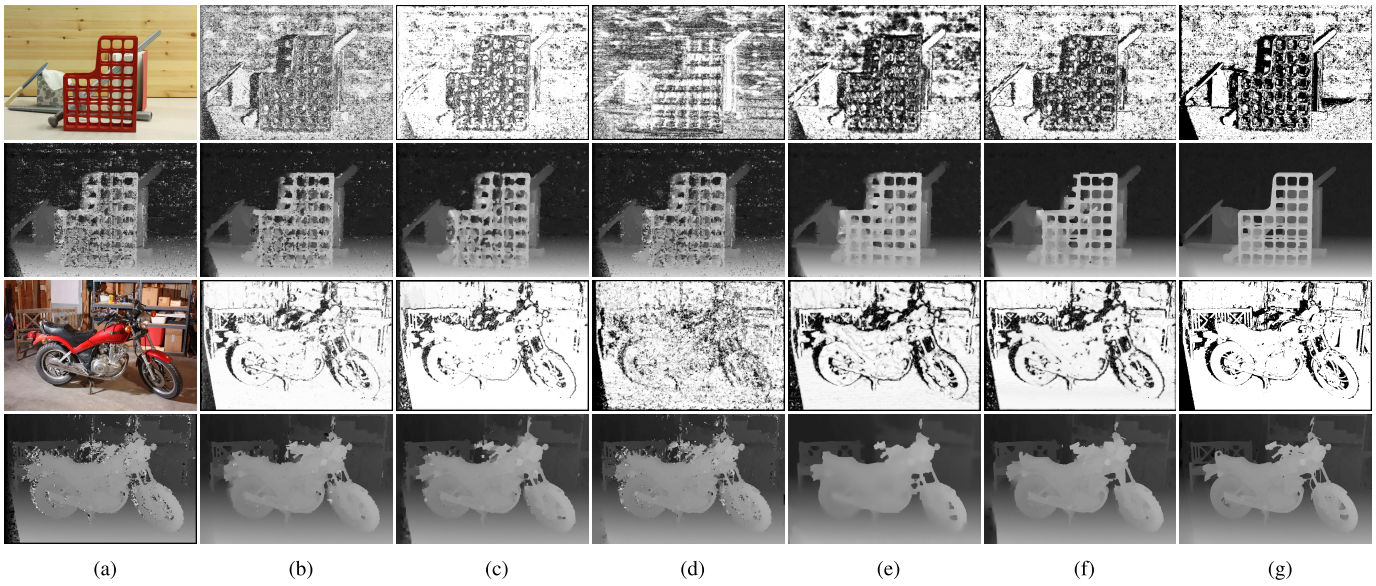
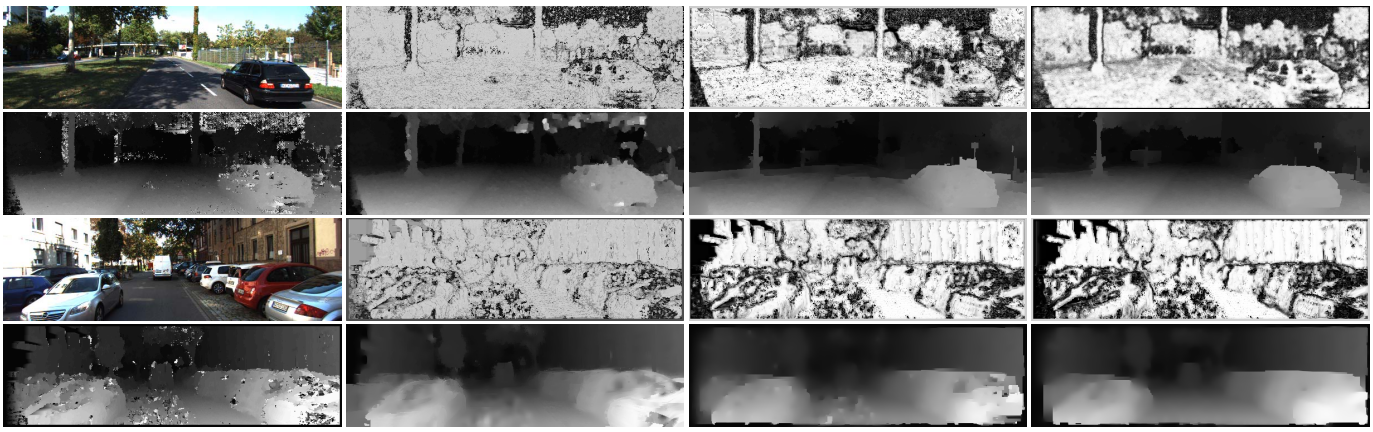


Fig. 14. The confidence and refined disparity maps on MID 2014 dataset [40] using census-SGM (first two rows), and MC-CNN (last two rows). (a) color images and initial disparity maps, refined disparity maps with corresponding confidence maps estimated by (b) Haeusler *et al.* [18], (c) Poggi and Mattoccia [21], (d) Shaked and Wolf [8], (e) Kim *et al.* (Conf) [29], (f) Ours (Conf), and (g) ground-truth confidence map.

propogation without additional post-processing to clearly show the performance gain achieved by the confidence measure. To evaluate the quantitative performance, we measured an average BMP for the MID 2006 [39], MID 2014 [40], and KITTI 2015 [41] datasets. Table IV show the BMP with thresholds of one and three pixels for MID 2006 [39], MID 2014 [40], and KITTI 2015 [41] dataset 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. The proposed method achieves the lowest BMP in all experiments.



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

Fig. 15. The confidence and refined disparity maps on KITTI 2015 dataset [41] using census-SGM (first two rows), and MC-CNN (last two rows). (a) color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Spyropoulos *et al.* [14], (c) LFN [28], and (d) Ours (Conf). Our method has shown consistently reliable performances on pole and car regions.

TABLE III

THE AVERAGE AUC AND MSE VALUES FOR MID 2006 [39], MID 2014 [40], AND KITTI 2015 [41] DATASET. THE AUC VALUE OF GROUND-TRUTH CONFIDENCE IS MEASURED AS 'LOWER BOUND'. THE RESULT WITH THE LOWEST AUC VALUE IN EACH EXPERIMENT IS HIGHLIGHTED

Datasets	MID 2006 [40]				MID 2014 [41]				KITTI 2015 [42]			
	Census-SGM		MC-CNN		Census-SGM		MC-CNN		Census-SGM		MC-CNN	
	AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE
Haeusler et al. [18]	0.0454	0.1892	0.0417	0.1777	0.0841	0.1965	0.0750	0.1756	0.0585	0.3295	0.0308	0.3218
Spyropoulos et al. [14]	0.0447	0.1451	0.0420	0.1478	0.0839	0.1681	0.0752	0.1377	0.0536	0.2783	0.0323	0.1976
Park and Yoon [15]	0.0438	0.1112	0.0426	0.1049	0.0802	0.1633	0.0734	0.1078	0.0527	0.2019	0.0303	0.1270
Poggi et al. [16]	0.0439	0.1148	0.0413	0.1025	0.0791	0.1326	0.0707	0.1046	0.0461	0.1628	0.0263	0.1488
Kim et al. [17]	0.0430	0.1219	0.0409	0.1354	0.0772	0.1403	0.0701	0.1266	0.0430	0.2356	0.0294	0.2330
CCNN [21]	0.0454	0.0890	0.0402	0.0953	0.0769	0.1129	0.0716	0.1014	0.0419	0.1088	0.0258	0.1254
PBCP [22]	0.0462	0.0983	0.0413	0.0909	0.0791	0.1204	0.0718	0.1090	0.0439	0.1355	0.0272	0.1532
Shaked et al. (Conf) [8]	0.0464	0.1138	0.0495	0.1049	0.0806	0.1063	0.0736	0.1014	0.0531	0.1067	0.0292	0.1358
LFN [28]	0.0416	0.0802	0.0393	0.0823	0.0752	0.1071	0.0692	0.0898	0.0405	0.0919	0.0253	0.1077
Kim et al. (Conf) [29]	0.0419	0.0855	0.0394	0.0828	0.0749	0.1035	0.0694	0.0912	0.0407	0.0927	0.0250	0.1108
Ours (Conf)	**0.0411**	**0.0781**	**0.0376**	**0.0716**	**0.0702**	**0.0823**	**0.0688**	**0.0754**	**0.0391**	**0.0628**	**0.0231**	**0.0990**
Lower bound	0.0340	-	0.0323	-	0.0569	-	0.0527	-	0.0348	-	0.0170	-

TABLE IV

THE BMP OF THE RESULTANT DISPARITY MAP ON MID 2006 [39], MID 2014 [40], AND KITTI 2015 [41] DATASET WITH CENSUS-BASED SGM. THE BAD PIXEL ERROR RATE OF REFINED DISPARITY MAPS USING GROUND-TRUTH CONFIDENCE IS MEASURED AS 'LOWER BOUND'. THE BMP IS MEASURED WITH ONE AND THREE PIXEL ERRORS. THE RESULT WITH THE LOWEST BMP IN EACH EXPERIMENT IS HIGHLIGHTED

Datasets	MID 2006 [40]				MID 2014 [41]				KITTI 2015 [42]			
	Census-SGM		MC-CNN		Census-SGM		MC-CNN		Census-SGM		MC-CNN	
	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px	>1px	>3px
Initial disparity	20.43	10.18	17.04	8.31	43.17	24.46	39.56	20.69	23.50	18.67	20.62	15.79
Haeusler et al. [18]	11.82	7.06	9.49	6.13	33.66	10.81	31.18	11.48	11.56	10.18	12.31	10.34
Spyropoulos et al. [14]	11.36	6.51	9.62	5.86	31.97	10.62	30.64	11.02	11.46	9.59	11.31	8.61
Park and Yoon [15]	10.98	6.03	9.33	5.72	29.16	10.15	28.31	10.71	10.37	9.18	11.28	8.12
Poggi et al. [16]	9.58	5.77	8.28	5.15	27.03	9.90	27.19	9.44	9.69	8.12	10.46	7.97
Kim et al. [17]	7.72	5.08	8.35	4.93	23.74	8.53	26.47	9.42	8.50	7.08	9.72	7.54
CCNN [21]	7.84	4.63	8.23	4.13	23.98	8.89	25.71	9.23	8.68	6.79	8.83	7.11
PBCP [22]	7.62	4.02	7.80	4.02	23.16	8.88	24.56	8.88	7.56	6.46	9.02	6.63
Shaked et al. (Conf) [8]	10.42	7.18	10.46	6.11	32.08	11.41	30.76	11.12	12.68	10.41	12.05	10.13
LFN [28]	7.27	3.86	7.18	3.82	22.65	8.62	23.17	8.38	7.28	6.21	8.52	6.49
Kim et al. (Conf) [29]	7.16	3.85	7.27	3.76	22.58	8.40	23.09	8.55	7.24	6.18	8.20	6.18
Ours (Conf)	**6.38**	**3.35**	**6.27**	**3.23**	**21.96**	**7.93**	**21.87**	**7.52**	**6.83**	**5.94**	**7.69**	**5.77**
Lower bound	3.45	1.97	4.02	1.56	20.53	5.27	17.61	4.54	4.57	3.51	3.59	3.31

Fig. 13, Fig. 14, Fig. 15, and Fig. 16 display the disparity maps refined with the confidence maps estimated from the existing handcrafted classifiers [14]–[17], CNN-based classifiers [8], [21], [22], [28], and our method, respectively. It was clearly shown that the erroneous matches are reliably removed using our confidence estimator. For the KITTI 2015 dataset [41] and Cityscapes dataset [59], erroneous disparities usually occur in textureless regions (e.g., sky and

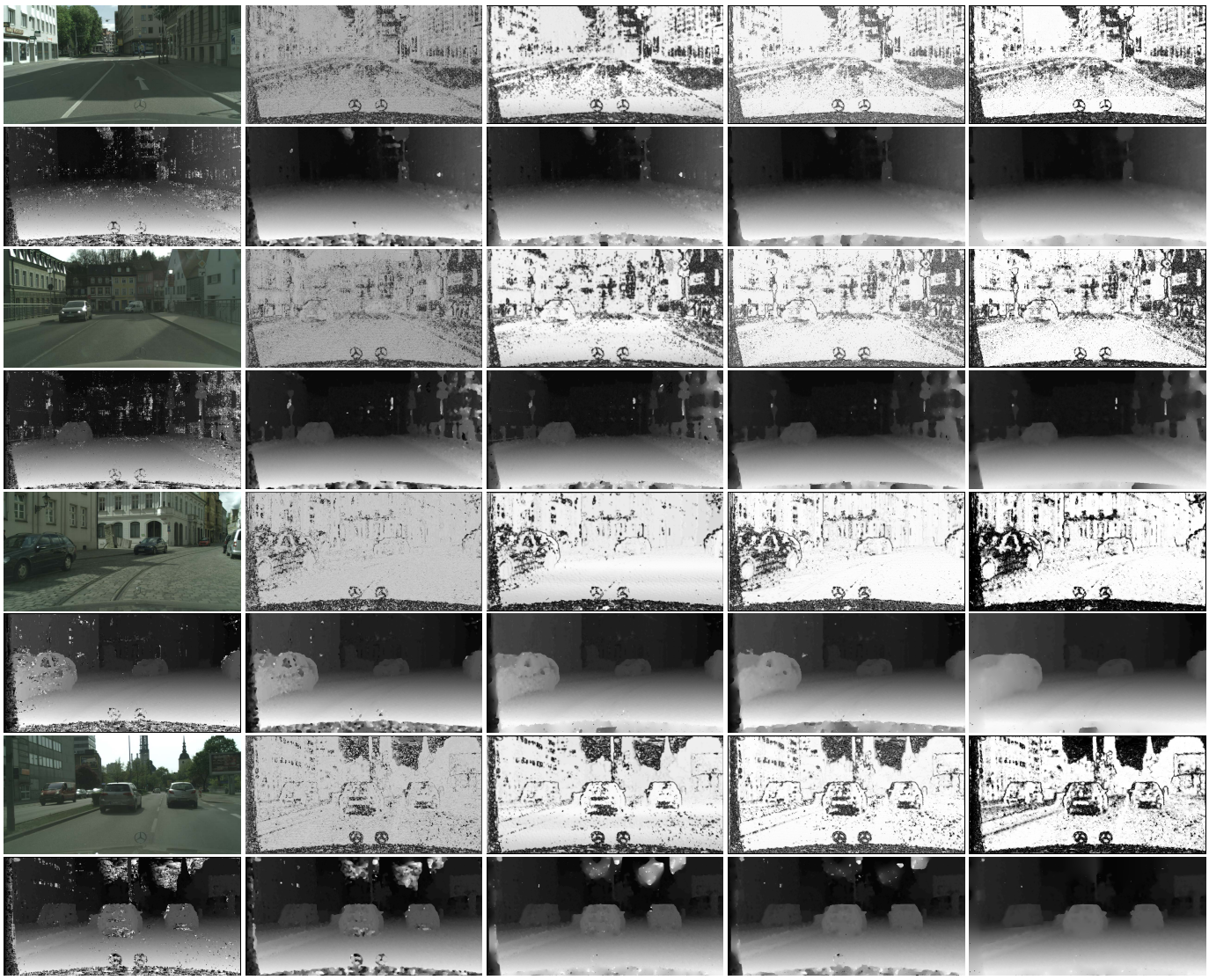


Fig. 16. The confidence and refined disparity maps on Cityscapes dataset [59] using census-SGM (1-4 rows), and MC-CNN (5-8 rows). (a) color images and initial disparity maps, refined disparity maps with confidence maps estimated by (b) Park and Yoon [15], (c) CCNN [21], (d) LFN [28], and (e) Ours (Conf). Our method has shown consistently reliable performances on pole and car regions.

road), as exemplified in Fig. 15 and Fig. 16. Conventional approaches [16], [21], [22], [28] show the limited performance for detecting incorrect pixels in 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.

VII. Conclusion

We presented a method that jointly estimates disparity and confidence through deep networks in an adversarial fashion. We formulated the generative cost aggregation and discriminative confidence estimation networks as two adversarial players. We proved that the discriminative confidence estimation networks not only generate the confidence map, but operate as the discrimator to boost the cost aggregation networks. In addition, the dynamic fusion module was presented to benefit from complementary information of matching cost, disparity, and color image in confidence estimation. Experimental results have shown that this method can improve the disparity and confidence estimation performance even in challenging real driving circumstances. A direction for further study is to examine how the proposed networks could be learned in a fully unsupervised fashion.

References

[1] 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.

[2] 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.

[3] 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.

[4] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D reconstruction in real-time," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 963–968.

[5] T. Cao, Z.-Y. Xiang, and J.-L. Liu, "Perception in disparity: An efficient navigation framework for autonomous vehicles with stereo cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2935–2948, Oct. 2015.

[6] K.-J. Yoon and I. So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

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

[8] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4641–4650.

[9] 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. Workshops (ICCV Workshops)*, Nov. 2011, pp. 467–474.

[10] 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, Nov. 2010.

[11] 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.

[12] R. Zabih and J. Woodfil, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Eur. Conf. Comput. Vis.*, May 1994, pp. 151–158.

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

[14] 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. (CVPR)*, Jun. 2014, pp. 1621–1628.

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

[16] 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. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 509–518.

[17] 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.

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

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 63, no. 4, pp. 5–32, 2001.

[20] A. Liaw and M. Wiener, "Classification and regression by random forest," *R Newslett.*, vol. 2, no. 3, pp. 18–22, 2002.

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

[22] A. Seki and M. Pollefeys, "Patch based confidence prediction for dense disparity map," in *Proc. Brit. Mach. Vis. Conf.*, vol. 10, Sep. 2016, pp. 1–4.

[23] S. Kim, D. Min, B. Ham, S. Kim, and K. Sohn, "Deep stereo confidence prediction for depth estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 992–996.

[24] M. Poggi and S. Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2452–2461.

[25] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5228–5237.

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

[27] 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, pp. 319–334.

[28] Z. Fu and M. Ardabilian Fard, "Learning confidence measures by multi-modal convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1321–1330.

[29] S. Kim, D. Min, S. Kim, and K. Sohn, "Unified confidence estimation networks for robust stereo matching," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1299–1313, Mar. 2019.

[30] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 2672–2680.

[31] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 649–666.

[32] R. Zhang *et al.*, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.

[33] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, p. 4.

[34] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. IEEE Conf. Learn. Represent.*, May 2016, pp. 1–16.

[36] C. Pu, R. Song, R. Tylecek, N. Li, and R. B. Fisher, "Sdf-GAN: Semi-supervised depth fusion with multi-scale adversarial networks," 2018, *arXiv:1803.06657*. [Online]. Available: http://arxiv.org/abs/1803.06657

[37] H. Huang, B. Huang, H. Lu, and H. Weng, "Stereo matching using conditional adversarial networks," in *Proc. Int. Conf. Neural Inf. Netw.*, 2017, pp. 124–132.

[38] 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.*, Jun. 2014, pp. 1725–1732.

[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. (GCPR)*, Sep. 2014, pp. 31–42.

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

[42] 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, Oct. 2004.

[43] P. Mordohai, "The self-aware matching measure for stereo," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1841–1848.

[44] 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.

[45] A. C. Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 300–308.

[46] M. Liang, X. Guo, H. Li, X. Wang, and Y. Song, "Unsupervised cross-spectral stereo matching by learning to synthesize," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 8706–8713.

[47] A. Kendall *et al.*, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[49] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Jun. 2015, pp. 2017–2025.

[50] 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.

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

[52] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1–11.

[53] A. Geiger, P. Lenz, and R. Urtasun. *Are We Ready for Autonomous Driving? The Kitti Vision Benchmark Suite*. Accessed: Dec. 15, 2018. [Online]. Available: http://www.cvlibs.net/datasets/kitti/

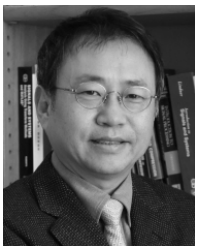
[54] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[55] 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.
[56] A. Vedaldi and K. Lnc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2015, pp. 689–692.
[57] *MC-CNN Github*. Accessed: Mar. 1, 2017. [Online]. Available: http://github.com/jzbontar/mc-cnn
[58] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 611–625.
[59] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.

Sunok Kim (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2014 and 2019, respectively. Since 2019, she has been a Post-Doctoral Researcher with the School of Electrical and Electronic Engineering, Yonsei University. Her current research interests include 3-D image processing and computer vision, in particular, stereo matching, depth super-resolution, and confidence estimation.

Dongbo Min (Senior Member, IEEE) 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 (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2012 and 2018, respectively. From 2018 to 2019, he was a Post-Doctoral Researcher with Yonsei University, Seoul. From 2019 to 2020, he was a Post-Doctoral Researcher with the School of Computer and Communication Sciences, École Polytechnique Féd érale de Lausanne (EPFL), Lausanne, Switzerland. Since 2020, he has been an Assistant Professor with the Department of Computer Science and Engineering, Korea University, Seoul. His current research interests include 2-D/3-D computer vision, computational photography, and machine learning.



Kwanghoon Sohn (Senior Member, IEEE) 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 3-D image processing and computer vision.