Stereo Confidence Estimation via Locally Adaptive Fusion and Knowledge Distillation

Sunok Kim, *Member, IEEE,* Seungryong Kim, *Member, IEEE,* Dongbo Min, *Senior Member, IEEE,* Pascal Frossard, *Fellow, IEEE,* Kwanghoon Sohn, *Senior Member, IEEE,*

Abstract—Stereo confidence estimation aims to estimate the reliability of the estimated disparity by stereo matching. Different from the previous methods that exploit the limited input modality, we present a novel method that estimates confidence map of an initial disparity by making full use of tri-modal input, including matching cost, disparity, and color image through deep networks. The proposed network, termed as Locally Adaptive Fusion Networks (LAF-Net), learns locally-varying attention and scale maps to fuse the tri-modal confidence features. Moreover, we propose a knowledge distillation framework to learn more compact confidence estimation networks as student networks. By transferring the knowledge from LAF-Net as teacher networks, the student networks that solely take as input a disparity can achieve comparable performance. To transfer more informative knowledge, we also propose a module to learn the locally-varying temperature in a softmax function. We further extend this framework to a multiview scenario. Experimental results show that LAF-Net and its variations outperform the state-of-the-art stereo confidence methods on various benchmarks.

Index Terms—Stereo matching, stereo confidence estimation, knowledge distillation, deep learning

1 Introduction

Stereo matching for reconstructing geometric configuration of a scene is one of the fundamental and essential problems in Computer Vision fields [1], [2], [3]. For decades, numerous methods have been proposed for this task by leveraging handcrafted [1], [4] and/or machine learning based [5], [6] techniques. However, because of its challenging elements such as reflective surfaces, textureless regions, repeated pattern regions, occlusions [7], [8], [9], and photometric deformations incurred by illumination and camera specification variations [10], [11], stereo matching still remains one of the unsolved problems. To alleviate these inherent challenges, most methods [12], [13], [14], [15], [16], [17] have adopted the confidence estimation step that detects unreliable disparities and refines them for improving the quality of stereo matching results.

Formally, the stereo confidence estimation pipeline involves first extracting the confidence features and then training the confidence classifiers using ground-truth confidences [12], [13], [18]. Conventionally, there exist several handcrafted confidence measures using different input modalities, such as matching cost, disparity, and color image [19], [20]. Since any single confidence measure cannot handle all failure cases in stereo matching, various combinations of hand-designed confidence measures extracted from the tri-modal input [12], [13], [14], [15], [21] have been used to learn shallow classifiers, such as random decision forest [22], [23]. Despite some performance improvement by the joint usage of the tri-modal input, they still show a limited performance due to their low discriminative power.

The most recent approaches have attempted to estimate the confidence by leveraging deep convolutional neural networks (CNNs) thanks to their high robustness and discriminative power [16], [17], [18], [24], demonstrating the substantial accuracy gain over the handcrafted approaches. However, unlike handcrafted approaches [12], [13], [21] that make full use of the tri-modal input, they have been formulated by partially using single- or bi-modal input, e.g., matching cost only [6], disparity only [18], [24], matching cost and disparity [16], [17], or disparity and color [25], [26]. Moreover, a simple concatenation technique [27] is commonly used to fuse tri-modal confidence features, disregarding that the fusion weights may vary for each pixel, depending on the attribute of confidence features.

In this paper, we propose novel confidence estimation networks, called Locally Adaptive Fusion Networks (LAF-Net), that utilize tri-modal input consisting of matching cost, disparity, and color image. The networks consist of confidence feature extraction networks, attention inference networks, scale inference networks, and recursive confidence refinement networks. In the attention inference networks, we fuse the tri-modal input adaptively with locally-varying attention maps to benefit from the joint usage of the tri-modal confidence features. In the scale inference networks, locally adaptive scale parameters are learned for all pixels, to extract the confidence features within locally optimal receptive fields. The output confidence is further refined through the recursive confidence refinement networks.

Even though LAF-Net provides improved performance in accuracy, it demands high computational burden that hinders its applicability on light-weight devices in real-world applications. To overcome this, we further present a light-weight confidence estimation model that achieves

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2021R1A2C2006703). The work of S. Kim was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2021R1C1C2005202).

- *S. Kim is with Korea Aerospace University, Goyang, Korea*
- *S. Kim is with Korea University, Seoul, Korea*
- *D. Min is with Ewha Womans University, Seoul, Korea*
- *P. Frossard is with EPFL, Lausanne, Switzerland*
- *K. Sohn is with Yonsei University, Seoul, Korea*

comparable performance to the LAF-Net through proposed knowledge distillation framework. By effectively transferring the knowledge from the LAF-Net as a teacher, the student networks that solely take as input a disparity can achieve comparable performance to the teacher that takes matching cost, disparity, and color image.

In addition, we extend the proposed methods into multiview stereo (MVS) setting. There exist methods to estimate the confidence in multiview images [28], [29], [30], [31], [32], where they used confidence to improve the output disparity, but all the methods leveraged limited information from reference view only, i.e., an initial disparity estimated from neighboring view, and did not consider multi-modal input of various information, such as matching cost or color image, from neighboring views. To have benefits from multiview images, we suggest a new multiview stereo confidence estimation network by learning the confidence features by leveraging geometric and photometric consistency. By considering such a network as teacher, we then apply knowledge distillation framework to learn student networks which can achieve improved performance gains in terms of both accuracy and complexity.

The proposed method is extensively evaluated through an ablation study and comparison with conventional handcrafted and CNNs-based methods on various benchmarks, including Middlebury 2006 [33], Middlebury 2014 [34], KITTI 2015 [35], and SUN3D [36].

This manuscript extends the conference version [37] of this work. It newly adds (1) the knowledge distillation framework for more efficient model; (2) an extension to multiview stereo confidence estimation; and (3) an extensive comparative study using various datasets.

2 Related Works

2.1 Stereo Confidence Estimation

For last decades, numerous approaches have been proposed to estimate reliable stereo confidence [12], [13], [14], [15], [16], [17], [18], [21], [24], [37]. An extensive study of handcrafted confidence measure have been studied by Park and Yoon [20]. The approaches with hand-designed confidence features have difficulty detecting unreliable pixels in challenging scenes. Recently, there have been attempts to estimate the confidence with deep CNNs [16], [17], [18], [24], [37]. In [38], a quantitative evaluation of confidence measure that use machine learning techniques has been performed. CCNN [18] is the first attempt to predict confidence through deep CNNs using left disparity. There have been various methods using single- or bi-modal input, such as left disparity [18], left and right disparity [24], matching cost and disparity [15], [16], disparity and color [25], [26]. Although these methods improved performance, they did not make full use of the tri-modal input, thus having limited performance.

In [17], they suggested multi-scale disparity feature extractor, while dilation convolution was applied in [25] to gain local contextualized information effectively. In [26], they proposed global confidence measure using encoder-decoder networks by looking at the whole image and disparity content. By using the output of global confidence, they proposed a local-global approach by fusing the local confidence, the global confidence, and disparity. All of these methods considered only fixed and pre-defined scale ranges and did not estimate a scale that varies for each pixel. On the other hand, the confidence refinement networks [39] were also developed, which can improve the accuracy of the estimated confidence map by leveraging a local consistency within the confidence map. There were no attempts to incorporate such designs within end-to-end networks.

On the other hand, there have been some approaches for estimating confidence for multiview stereo (MVS) [28], [30], [31], [32]. In [28], the authors suggest the cross-view confidence measure and spatial confidence measure to improve the multiview depth estimation. Assuming that the uncertainty is related to the distribution of matching cost, some methods [31], [32] jointly learn the depth and confidence by minimizing the negative log likelihood. In [30], they estimate MVS confidence with the multi-modal data consisting of normal map, depth map, and color image. All of these methods predict confidence map using only data of reference view and not using the additional cues from neighboring views.

2.2 Knowledge Distillation

Knowledge distillation is developed to transfer the rich information of a teacher model to a student model for improving the performance of the student model [40]. In this framework, student networks are trained using the softmax outputs of a teacher networks as soft labels, which enables learning more informative features of the student networks. Inspired by [40], many variants have been introduced such as feature distillation [41], modality distillation [42], and self distillation [43]. Compared to conventional knowledge distillation [40] or feature distillation [44], [45], which use the same input in both training and testing procedure, modality distillation [42] has benefits from extra information in teacher networks. In [42], video action recognition using modality distillation have been proposed by taking RGB and depth video as input in training and relying on RGB only in testing. Similarly, learning with privileged information [46], [47] or side information [48] leverages extra information at training but with no accessibility to it at testing. Recently, unified frameworks of distillation and privileged information have been introduced [49], [50], named generalized distillation. In [50], they proposed graph distillation which incorporates privileged information from a multi-modal dataset in the source domain, and improves the learning performance in the target domain where training data and modalities are scarce.

Our method belongs to generalized distillation, since we train teacher networks with multi-modal dataset consisting of matching cost, disparity, and color image, which can be considered as privileged information, but only use single disparity to learn student networks.

3 Locally Adaptive Fusion (LAF) Networks

3.1 Overview

Most existing methods for stereo matching, even with deep CNNs [1], [3], [5], [43], cannot provide fully reliable results due to inherent challenges of the task, and thus, several approaches [12], [13], [15], [16], [24] presented an additional

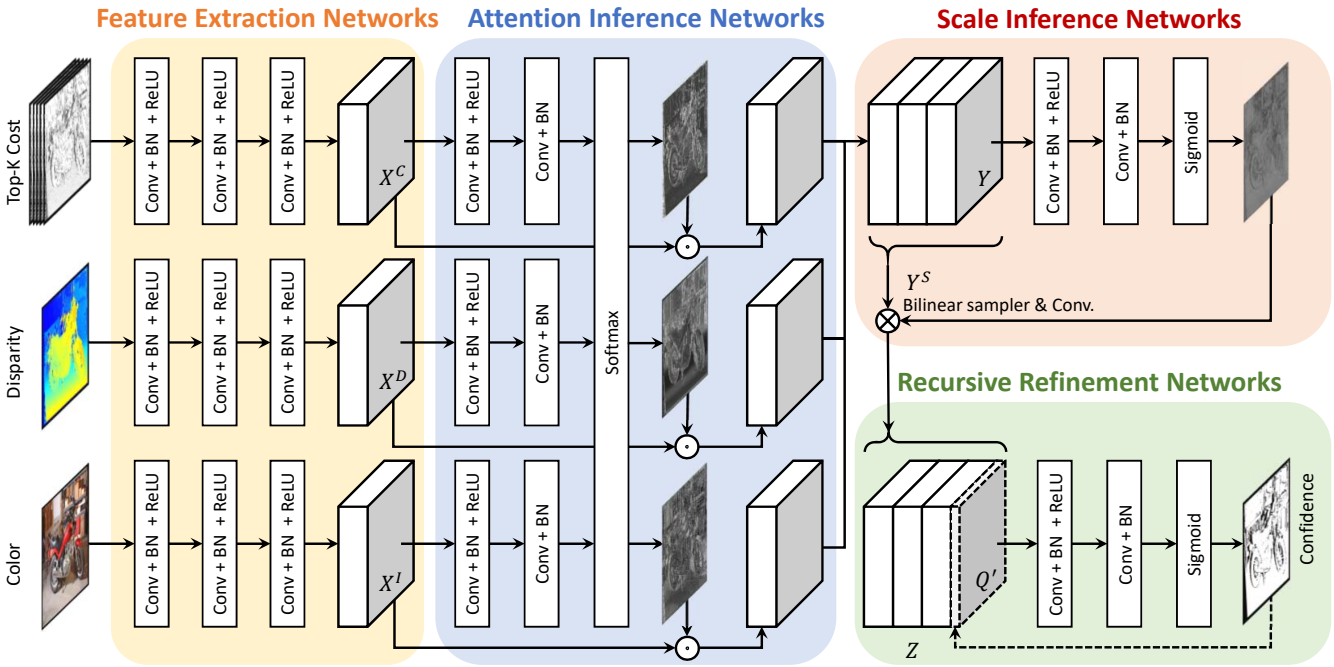


Fig. 1. **Network configuration.** LAF-Net consists of four sub-networks, including feature extraction networks, attention inference networks, scale inference network, and recursive refinement networks. Given matching cost, disparity, and image as input, it outputs confidence of the disparity.

module to predict a confidence of disparity. By leveraging the confidence, they refine the initial disparity through subsequent disparity refinement pipeline.

Formally, let us define stereo image pair as I^l and I^r, respectively. The objective of stereo matching is to estimate a disparity D_i between the pair that is defined for pixel $i = [i_{\mathrm{x}}, i_{\mathrm{y}}]^T$. To this end, the matching costs $C_{i,d}$ between I_i^l and $I_{i-[d,0]^T}^r$ among candidates $d \in \{1, ..., d_{\max}\}$ are first measured, and then aggregated and optimized. To estimate the confidence Q of disparity D, unlike conventional handcrafted approaches [12], [13], [21] that make full use of the tri-modal input, such as matching cost C, disparity D, and image I, recent CNNs-based methods partially used single- or bi-modal input [6], [16], [17], [18], [24], [25], [26], thus providing limited performance under challenging environments.

In this work, we design a novel network architecture that estimates the confidence by fully exploiting matching cost, disparity, and color. The overall networks consist of four sub-networks, including *feature extraction* networks, *attention inference* networks, *scale inference* networks, and *recursive refinement* networks, as illustrated in Fig. 1. In feature extraction networks, confidence features are independently extracted from the tri-modal input. The intermediate features are then fed to learn locally-varying attention maps in attention inference networks, which are used to adaptively fuse the tri-modal features, unlike existing approaches [16], [17], [25], [26] that use a simple concatenation. Then, locally-varying scale fields are learned for extracting confidence features within geometrically-aligned receptive fields through scale inference networks, different from conventional approaches [17], [18], [24] with a fixed-size convolution. Finally, the confidence is progressively refined in recursive confidence refinement networks to enforce a spatial context and local consistency inspired by [15], [39]. In the following, we describe each module in details.

3.2 Feature Extraction Networks

Due to their heterogeneous attributes, direct concatenation of the tri-modal input does not guarantee an optimal fusion [25]. Alternatively, we design the feature extraction networks to extract the heterogeneous tri-modal features denoted as X^C, X^D, and X^I from matching cost C, disparity D, and color image[1] I by feed-forward processes such that $X^C = \mathcal{F}(C; W^C)$, $X^D = \mathcal{F}(D; W^D)$, and $X^I = \mathcal{F}(I^l; W^I)$ with parameters W^C, W^D, and W^I, respectively. The network parameters for each network are separately learned to encode the different characteristics of the tri-modal input.

The absolute value of matching cost may vary depending on the search range of stereo image pairs and stereo matching methods used. In addition, its distribution is often non-discriminative, as mentioned in [24], [25]. To alleviate these limitations, the matching cost is transformed into a top-K matching probability[2] as in [16], [17], which enables achieving the search range-invariance.

3.3 Attention Inference Networks

Some methods [16], [17], [25], [26] first extract the bi-modal confidence features and then concatenate them. However, such a simple concatenation at inference often fails to perform an optimal fusion. To alleviate this limitation, inspired by [51], we build the attention inference networks for inferring an optimal fusion weight between the tri-modal features, i.e., X^C, X^D, and X^I. The locally-varying attention for each modality is defined as A_i^C, A_i^D, and A_i^I at pixel i for matching cost, disparity, and color image

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

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

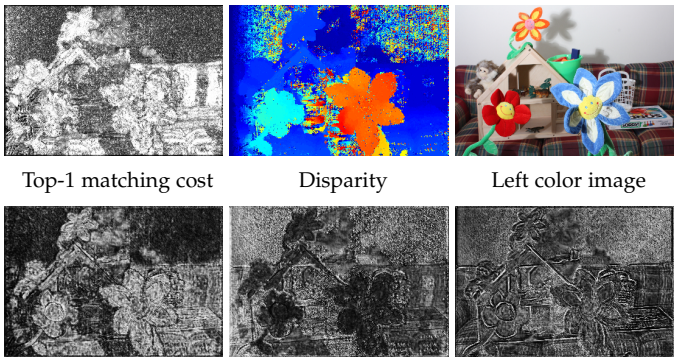


Fig. 2. **Visualization of learned attention maps:** top-1 matching cost, disparity, left color image, and the attention maps for matching cost, disparity, and color, respectively.

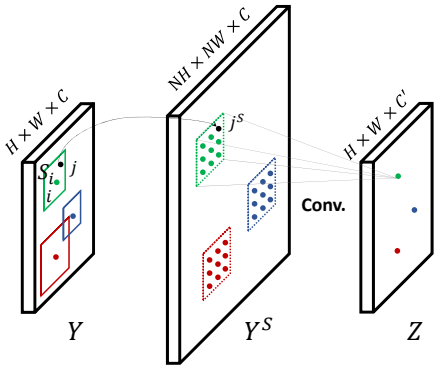


Fig. 3. **Illustration of bilinear sampler in scale inference networks.** For each pixel i, the feature Y can be warped as enlarged size feature Y^S. With the stride, the neighbors j^S are convolved as Z.

such that $A_i^C = \mathcal{F}(X_i^C; W_A^C)$, $A_i^D = \mathcal{F}(X_i^D; W_A^D)$, and $A_i^I = \mathcal{F}(X_i^I; W_A^I)$ with the parameters W_A^C, W_A^D, and W_A^I, respectively. These attentions then undergo a softmax function to make the sum of attentions for each pixel to be 1, i.e., $\sum_{x \in C,D,I}(A_i^x) = 1$. Note that the attention inference network parameters for each modality (i.e., W_A^C, W_A^D, and W_A^I) are not shared but independently learned depending on their attributes.

The learned attentions are then applied to the tri-modal features, X^C, X^D, and X^I, to fuse them as

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

where $\Pi(\cdot)$ is a concatenation operator and $\odot$ is an element-wise multiplication operator. Note that unlike methods [17], [25], [26] using the fixed fusion weights, the attentions, A^C, A^D, and A^I, are estimated conditioned on input and vary locally, thus enabling the adaptive fusion more effectively.

The visualization of attention maps for different input modalities is exemplified in Fig. 2. The attention of matching cost A^C is high for pixels having high matching probability. On the other hand, the attention of disparity A^D has high value in noisy regions, indicating informative features can be extracted from the different disparity assignments, as considered similar to VAR or MDD [21] in handcrafted features. In color image, the attentions near image boundary A^I are high, which indicates an image texture can give a useful cue to estimate confidence. By adaptively weighting the features with these attention maps, we can obtain more discriminative features.

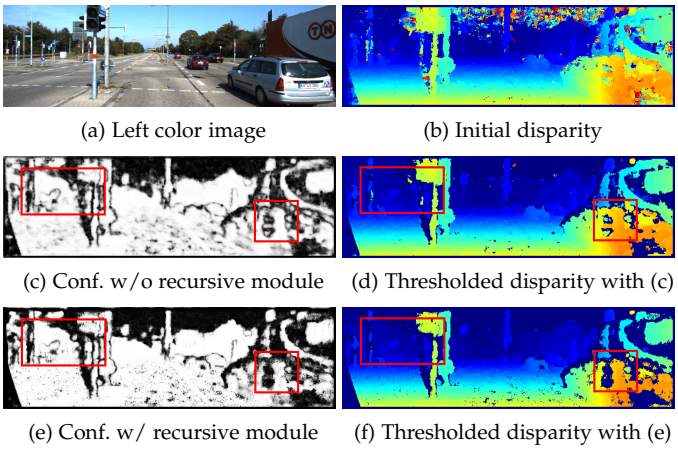


Fig. 4. **Effectiveness of recursive refinement networks:** (a) left color image, (b) initial disparity, (c) estimated confidence map without recursive module, (d) thresholded disparity with (c), (e) estimated confidence map with recursive module, (f) thresholded disparity with (e). The mismatched pixels in the red boxes are reliably detected with the proposed recursive refinement networks.

3.4 Scale Inference Networks

The optimal receptive fields for confidence features vary at each pixel. In order to encode features of different scales, some approaches [13], [17], [25], [26] have been proposed, but they consider only fixed and pre-defined scales and do not estimate scales that vary at each pixel. To determine such optimal receptive fields, we present the scale inference networks that learn locally-varying scale fields. It first infers the scale fields through subsequent convolutions such that $S_i = \mathcal{F}(Y_i; W^S)$ with parameters W^S. With these scale fields S_i, the intermediate receptive fields are warped through an image sampling on a parameterized grid, similar to spatial transformer networks (STNs) [52].

However, a spatially-varying parameterized sampling grid cannot be directly realized with the original STNs [52] that is designed for a global geometric field. To deal with locally-varying scale fields, we first build a locally-varying sampling grid for $N \times N$ neighbors $\mathcal{N}_i$ independently, and then warp the convolutional activation for each sampling grid as used in [53], [54]. Concretely, the locally-varying sampling grid $j^S = [j_{\mathbf{x}}^S, j_{\mathbf{y}}^S]^T$ is defined such that

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

for pixel i and their neighbors $j \in \mathcal{N}_i$ within receptive fields on the regular grid. For each grid sample j^S, receptive fields for convolution layers are warped through the bilinear sampler [52] independently such that

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

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

3.5 Recursive Refinement Networks

So far, we introduce our networks that fuse tri-modal confidence features through the attention and scale inference

networks. On the feature Z_i, we formulate the confidence prediction networks to estimate the confidence Q_i such that $Q_i = \mathcal{F}(Z_i; W^P)$ with the parameters W^P. The iterative refinement procedure of output confidence can improve the confidence estimation accuracy as studied in the hand-crafted approach using joint filtering [15] and CNNs-based approach [39]. Inspired by this, we propose the recursive refinement networks, where the previously estimated confidence serves as a guidance of the current estimation. To realize this recursive module, we formulate the networks to estimate confidence such that $Q_i^t = \mathcal{F}(Z_i, Q_i^{t-1}; W^P)$ where Q_i^t and Q_i^{t-1} are the estimated confidences at t^{th} and $(t-1)^{th}$ iteration, respectively. The initial confidence Q_i^0 is defined as zeros. As evolving the iterations, the accuracy is improved gradually and the final map is obtained as $Q^{t_{\max}}$ with the maximum number of iterations. The effectiveness of the recursive refinement networks is shown in Fig. 4. With the recursive module, the ability to predict mismatched pixels is improved.

3.6 Implementation Details

The feature extraction networks consist of 3 convolution layers with 3×3 kernels producing 64 feature channel, followed by batch normalization (BN) and rectified linear units (ReLU). The attention learning networks consist of 2 convolution layers with 3×3 kernels. The first convolution layer produces 64 channel feature, followed by BN and ReLU, and the second convolution layer produces 1 channel feature followed by only BN. In addition, the scale inference networks consist of 2 convolution layers with 3×3 kernels. The first convolution layer produces 64 channel feature, followed by BN and ReLU, and the second convolution layer produces 1 channel feature followed by only BN. The output passes through the sigmoid layer to generate the scale parameter for each pixel. The recursive refinement networks consist of 2 convolution layers and final sigmoid layer similar to the scale learning networks.

The overall networks are learned with the cross-entropy loss function [6], [17] as

$$\mathcal{L} = \mathcal{H}(Q', Q^*), \tag{4}$$

where Q' is estimated confidence and Q^* is ground-truth.

4 Distillation of LAF-Net

4.1 Motivation

Even though LAF-Net described above has been established as state-of-the-art in accuracy, as it will be shown in experiments Fig. 9, it demands high memory burden and complexity. To overcome this, we present a knowledge distillation framework to learn a simpler model but having competitive performance in comparison to LAF-Net. While the teacher networks, namely LAF-Net, take multi-modal inputs consisting of matching cost, disparity, and color image, we design the student networks to use only a disparity as input. In addition, we propose temperature inference networks which learn locally-varying temperature to transfer more informative soft label from teacher to student networks.

4.2 LAF-Net and Distilled LAF-Net

In order to train a simpler model as the student networks, named S-CCNN, we leverage an architecture similar to CCNN [18], which is one of the simplest methods among existing CNNs-based confidence estimators. It takes a single disparity only as input, and consists of only two stack of 3×3 convolution layers with 64 channel features followed by ReLU. Specifcially, we take LAF-Net as the teacher and S-CCNN as the student. We call such distilled S-CCNN networks as distilled LAF-Net (DLAF-Net).

We define final estimated confidence maps of LAF-Net and DLAF-Net as $Q_i^{\mathcal{T}}$ and $Q_i^{\mathcal{S}}$, respectively. For the knowledge distillation, we first generate a soft label S_i using a soft prediction output Q_i from the teacher networks such as

$$S_i = \mathrm{softmax}(Q_i/T), \tag{5}$$

where T is temperature to soften the output signals from teacher networks to provide more information to student networks as introduced in [40]. Existing knowledge distillation methods use a fixed temperature T, which is normally set to 1 [40], [44]. However, the *fixed (and global)* temperature is limited to transfer the meaningful information and optimal temperature may be varying for each pixel.

For a more effective knowledge distillation, we introduce temperature inference networks which learn *locally-varying* temperature T_i for each pixel i. The intuition behind this is that if the teacher generates unreliable confidences at some pixels, which may hinder the performance of student, its effects should be minimized by flattening Eq. (5), which is getting close to 0.5, where the student learns about 0.5 probability rather than learning wrong directions. To learn such optimal temperatures that can be locally varied, we design the temperature inference networks. In specific, we build the temperature inference networks by designing an extra branch from the final feature from the teacher network, consisting of 2 convolution layers with 3×3 kernels. The first convolution layer produces 64 channel feature, followed by BN and ReLU, and the second convolution layer produces 1 channel output, which is locally varying temperature T_i followed by BN. Note that we did not give any constraints on the output, so no activation function is used. The temperature inference networks were learned end-to-end with the student networks.

In our framework, soft label S_i in softmax function is defined such that

$$S_i = \mathrm{softmax}(Q_i^{\mathcal{T}}/T_i). \tag{6}$$

Then the distillation loss function is defined with cross-entropy loss:

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{H}(Q^{\mathcal{S}}, S). \tag{7}$$

The effectiveness of the temperature inference networks is exemplified in Fig. 6. The temperature T_i inferred by the proposed temperature inference networks has high value in the regions where the confidence is incorrectly estimated, which can soften the softmax output in Eq. (6). The effect of output confidence values of teacher networks can be reduced at these regions, which provides more informative knowledge to learn the student networks, while the soft label with fixed T simply normalizes the output confidence.

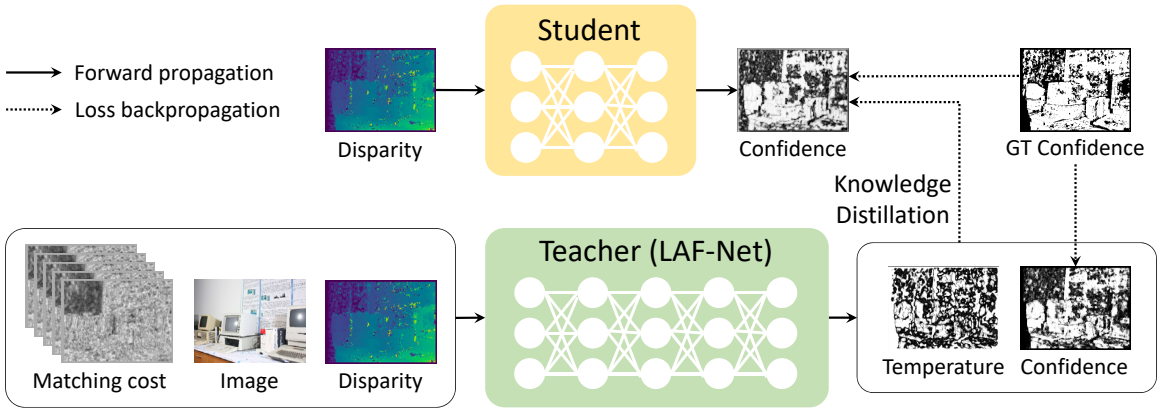


Fig. 5. **Network configuration in knowledge distilation framework.** The teacher networks (i.e., LAF-Net) take tri-modal input, while the student networks (i.e., DLAF-Net) take as input a disparity only. To train the student networks by transferring the knowledge of the teacher, we use the knowledge distilation framework with a learned temperature to boost the performance.

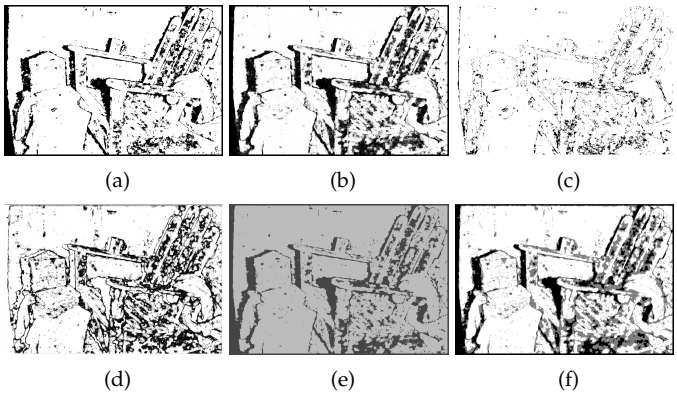


Fig. 6. **Effectiveness of temperature inference networks:** (a) ground truth confidence, (b) estimated confidence with teacher networks, namely LAF-Net, (c) difference between ground truth confidence (a) and estimated confidence (b), (d) temperature T inferred by the temperature inference networks, (e) soft label with fixed $T = 1$, and (f) soft label with T in (d). By using locally-varying learned temperatures, the effects of erroneous outputs of teach can be reduced, which boosts the distillation performance.

5 Generalization to Multiview Scenario

5.1 Overview

So far, we have introduced stereo confidence estimation methods in the two-view stereo setting. In this section, we extend the proposed methods to multiview stereo (MVS) scenario.

Given a reference image I and neighboring views I^n, with $n = 1, 2, ..., N$, the objective of MVS is to estimate a depth map D by comparing the points from the reference to neighbors. Similarly to stereo setting, most existing methods for MVS [29], [55], [56], [57], [58] cannot provide fully reliable results. We thus attempt to find a confidence map Q for D, which can be used in the depth refinement pipeline, similar to two-view setting [15], [16]. To realize this, we first generalize the LAF-Net for MVS confidence estimation, called multiview LAF-Net (M-LAF-Net), that infers the confidence by fully exploiting multi-modal data consisting of reference image I, its associated neighboring images I^n, a reference depth map D, its associated neighboring depth maps D^n, and matching cost C^n. We also present multiview DLAF-Net (M-DLAF-Net) through knowledge distillation framework. Similarly to knowledge distillation framework in two-view stereo, we take multi-modal data as input in the teacher networks while reference depth map is solely used in the student networks.

5.2 M-LAF-Net as Teacher Networks

Different from original LAF-Net, which uses matching cost, disparity, and color image for reference image, the proposed M-LAF-Net can access information from N-view images and depth maps in MVS setting. To leverage these, geometric and photometric consistency can be cues to detect correct estimation. To the best of our knowledge, this is first attempt to encode these consistencies to predict MVS confidence in deep neural networks.

5.2.1 *Encoding Geometric Consistency*

Our first key idea is that the each estimated depth map D^n obtained by matching the reference image I and its neighboring image I^n should be geometrically consistent after warped to the reference view [55] as shown in Fig. 7(a). In addition, since D^n is computed by the matching cost C^n, we take sum of each matching cost and multiple set of depth map as input to encode geometric consistency. As shown in Fig. 8, the concatenation across channel axis of neighboring depth maps can be used as input for extracting confidence features encoding geometric consistency between depth maps.

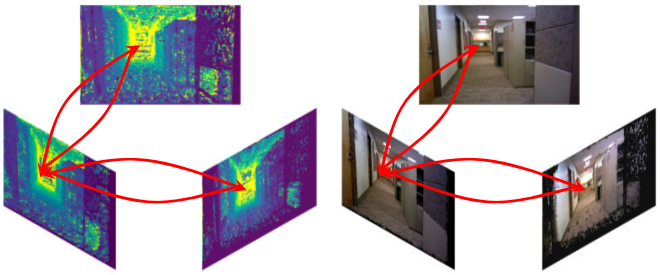
5.2.2 *Encoding Photometric Consistency*

Inspired by [56], we further leverage photometric consistency, where the accurate geometry prediction for a point should yield consistent predictions when projected onto other views as shown in Fig. 7(b). For a pair of views (I, I^n), where intrinsic and extrinsic (K, R^n) parameters are known, the estimated depth map D enables inverse-warping the novel view to the reference geometry. Specifically, for a pixel u in the reference image I, we can obtain its coordinate in the novel view with the warping function as follows:

$$\hat{u} = KR^n(D(u) \cdot K^{-1}u). \tag{8}$$

Then warped image $\hat{I}^n_s$ can be obtained such as

$$\hat{I}^n(u) = I^n(\hat{u}). \tag{9}$$



(a) Geometric consistency (b) Photometric consistency

Fig. 7. **Visualization of geometric and photometric consistency.** The depth map obtained by reference image and different views should be consistent as in (a). The warped image using depth map should be consistent as in (b).

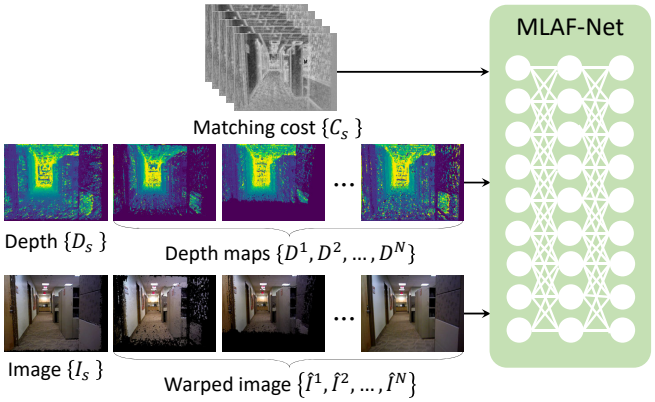


Fig. 8. **Illustration of M-LAF-Net.** Compared to LAF-Net, which takes the left disparity and image as inputs, we use the concatenation of multiple depth maps and warped images as inputs in M-LAF-Net.

Similar to geometric consistency, we can encode photometric consistency between warped images through a network that takes as input concatenation of the set of warped images, i.e., $\{\hat{I}^1, \hat{I}^2, ..., \hat{I}^N\}$.

5.2.3 Network Configuration

To summarize, we extend LAF-Net with additional inputs consisting of aforementioned multiple depth maps D^n and warped color images $\hat{I}^n$. In order to effectively encode the geometric and photometric consistency, we take the concatenation of depth images $\{D^1, D^2, ..., D^N\}$ and the concatenation of warped color images $\{\hat{I}^1, \hat{I}^2, ..., \hat{I}^N\}$ as input. The disparity and image branches of original LAF-Net is substituted with $\{D, D^1, D^2, ..., D_s^N\}$ and $\{I, \hat{I}^1, \hat{I}^2, ..., \hat{I}^N\}$. Similarly to original LAF-Net, the proposed M-LAF-Net yields high confidence estimation accuracy, but demands extremely high computational burden. Moreover, in MVS scenario, required memory may be much larger than stereo confidence estimation due to multiple set of inputs.

5.3 M-DLAF-Net as Student Networks

To benefit from the knowledge distillation proposed in Sec. 4, similarly to two-view stereo confidence estimation, we deploy an architecture similar to CCNN [18] as student networks which take a reference depth map as input, called multiview DLAF-Net (M-DLAF-Net). Also, we use confidence estimation networks in CLD-MVS [28] only CNNs-based parts to learn the networks. The entire networks can be learned with the same loss function in Sec. 4.

6 Experimental Results

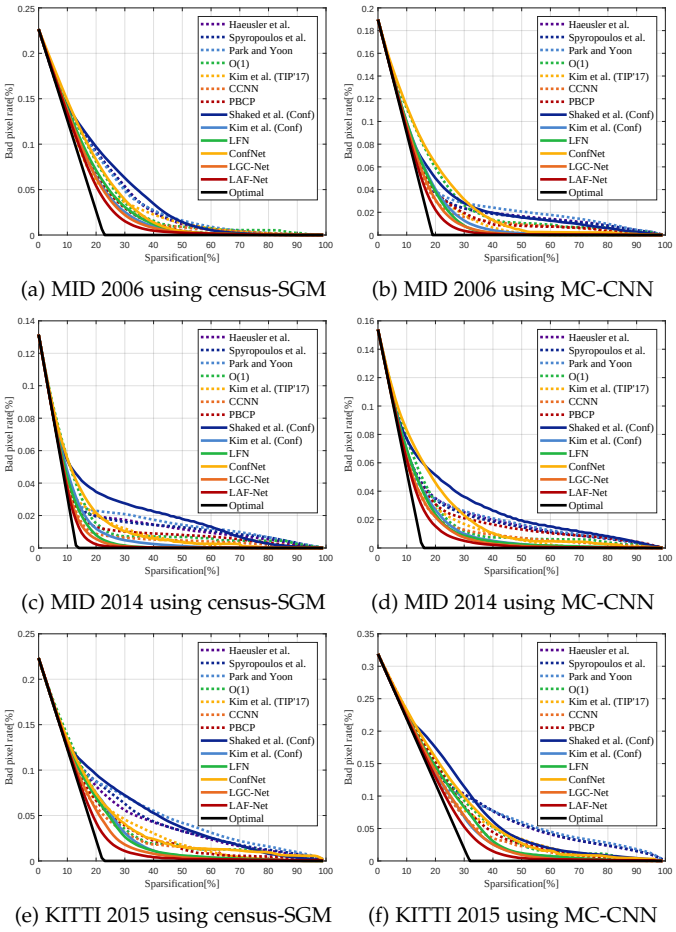
6.1 Experimental Settings

The proposed method was simulated on a Intel i7 CPU with NVIDIA RTX3090 GPU. We make use of the stochastic gradient descent with momentum, and set the learning rate to 1×10^{-6} and the batch size to 16. To compute a raw matching cost, we used a census transform with a 5×5 local window and MC-CNN [5], 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 [13]. For computing the MC-CNN, 'KITTI 2012 fast network' was used, provided at the author's website [59]. We set σ as 100 and 0.05 for census-SGM and MC-CNN, respectively, as in [17]. Finally, as in [60], we prove that our framework is effective also at improving confidences estimated for a deep stereo network such as GA-Net [43]. We trained our networks for two different datasets, i.e., MPI Sintel dataset [61] and KITTI 2012 dataset [35], to evaluate the performance across different database characteristic. The test datasets are Middlebury 2006 (MID 2006) [33], Middlebury 2014 (MID 2014) [34], and KITTI 2015 dataset [35]. We used the half-sized KITTI database (608×184). The ground-truth confidence maps are obtained by thresholding an absolute difference between estimated disparity and ground-truth disparity to 1. As in previous literature [18], [37], we set threshold 1 for MID 2006 and half-sized KITTI database and threshold 3 for MID 2014 and full-sized KITTI database, respectively. For the number of iteration in the recursive confidence estimation networks, we set $t_{\max}$ to 3. Our code is available at project page: https://github.com/seungryong/LAF/.

In the following, we evaluated the proposed method in comparison to conventional handcrafted approaches, such as Haeusler et al. [21], Spyropoulos et al. [12], Park and Yoon [13], Poggi and Mattoccia [14], Kim et al. [15]. Several CNNs-based approaches using single- or bi-modal input are also compared, where using disparity only, such as Poggi and Mattoccia (CCNN) [18], Seki and Pollefey (PBCP) [24], matching cost only, such as Shaked et al. [6], both disparity and matching cost, such as Kim et al. [17], and both color and disparity, such as Fu et al. (LFN) [25] and the global measures of Tosi et al. (ConfNet) [26] and local and global measures (LGC-Net) [26]. We obtained the results of [13], [15], and [17] by using the author-provided code, while the results of [21], [12], [24], [6], and [25] were obtained by our implementation. We re-implemented methods of [14], [18], and [26] based on author-provided code.

To evaluate the performance of confidence estimation quantitatively, we used the sparsification curve and its area under curve (AUC) as used in [12], [13], [17], [21], [24]. 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 estimation errors. For the higher accuracy of the confidence measure, AUC value is lower and the optimal AUC is measured using ground-truth confidence.

For MVS, we used SUN3D dataset [36]. This database consists of 415 sequences captured for 254 different spaces. We selected 36 sequences which contain more than 500

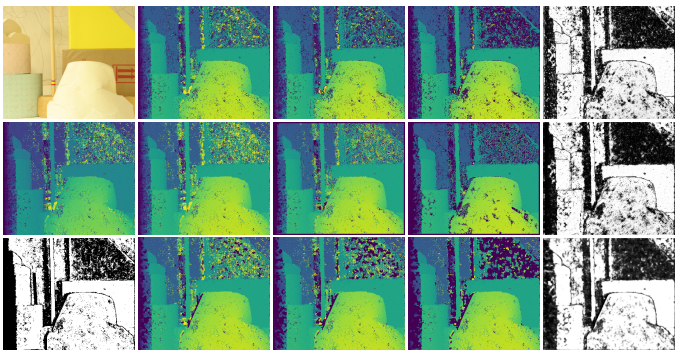


(a) MID 2006 using census-SGM (b) MID 2006 using MC-CNN

(c) MID 2014 using census-SGM (d) MID 2014 using MC-CNN

(e) KITTI 2015 using census-SGM (f) KITTI 2015 using MC-CNN

Fig. 9. **Sparsification curves of selected images for MID 2006 [33], MID 2014 [34], and KITTI 2015 dataset [35] using (a), (c), (e) census-SGM and (b), (d), (f) MC-CNN.** The sparsification curve for the ground-truth confidence map is described as 'optimal'.



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

Fig. 10. **Effects of specification according to confidence:** (a) color image, initial disparity map, and ground-truth confidence, initial disparity maps by gradually removing a percentage of pixels with lowest confidence by (top to bottom) CCNN [18], DLAF-Net, and LAF-Net, with (b) 5%, (c) 10%, (d) 20%, and (e) estimated confidence maps.

images for each sequence to deal with multi-view scenario and divided the train and test set into 30 and 6 sequences, respectively. In each sequence, we constructed 5 set of each having one reference image and four neighboring views. We will release the selected sequences for fair comparison. We used CCNN [18], S-CCNN as student networks and M-LAF-Net as teacher networks, comparing to the CNNs-based parts of CLD-MVS [28], LAF-Net, and DLAF-Net.

6.2 Stereo Confidence Estimation Analysis

In order to measure the performance of the confidence estimator in comparison to other methods, we compared the average AUC values of our method with conventional learning-based approaches using handcrafted confidence measures [12], [13], [14], [15], [21] and CNN-based methods [18], [24], [25], [26]. For fair comparison, we also evaluated the confidence estimation performance only for [6], [17], i.e., Shaked et al. (Conf) [6] and Kim et al. (Conf) [17].

Sparsification curves for Middlebury 2006 dataset [33], Middlebury 2014 dataset [34], and KITTI 2015 dataset [35] with census-based SGM and MC-CNN are shown in Fig. 9. Fig. 10 shows how, by removing the least confident pixels and selecting the most confident one, confidence estimators (including ours) are able to select a subset of pixels containing no outliers. The results have shown that the proposed confidence estimator exhibits a better performance than both conventional handcrafted approaches and CNN-based approaches. The average AUC with census-based SGM and MC-CNN for Middlebury 2006 (MID 2006), Middlebury 2014 (MID 2014), and KITTI 2015 datasets were summarized in Table 1. The handcrafted approaches showed inferior performance than the proposed method due to low discriminative power. CNN-based methods [6], [18], [24], [25] have improved confidence estimation performance compared to existing handcrafted methods such as [12], [13], [14], [15], [21], but they are still inferior to our method as they rely on single- [6], [18] or bi-modal [17], [24], [25], [26] input rather than tri-modal input. Also, the locally varying fusion weights improved the confidence estimation performance compared to methods [17], [25], [26] that use simple concatenation technique. The estimated confidence maps are shown in Fig. 11 and Fig. 12. Especially, the performance of DLAF-Net shows the competitive results of LGC-Net [26], which combines global and local confidence estimator, showing the effectiveness of proposed framework. The estimated confidences are shown in Fig. 13 and Fig. 15.

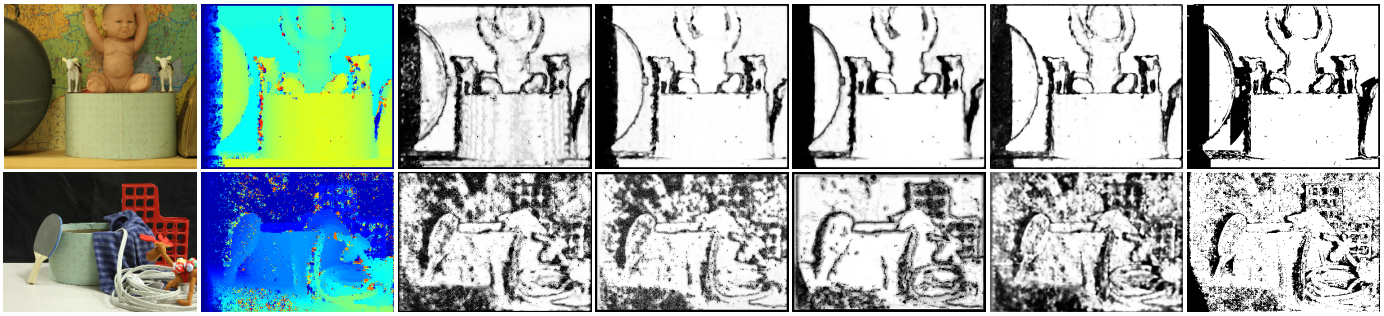
In addition, we measure the performance of our method with a modern deep network-based stereo matching algorithm, i.e., GA-Net [43]. Fig. 15 visualizes the qualitative results and Table 2 shows the quantitative results. Given the extremely regular structure of the estimated disparity maps by GA-Net [43], the cost volume becomes a crucial cue to properly estimate the confidence. Therefore our LAF-Net has shown the state-of-the-art performance on this experiment. DLAF-Net shows competitive performance although its complexity is much lower than LAF-Net.

6.3 MVS Confidence Estimation Analysis

The average AUC with MC-CNN for SUN3D dataset were summarized in Table 3. By leveraging the geometric and photometric consistency, M-LAF-Net have achieved the best results. Here, we note that the lower bound of the propposed confidence estimators, M-DLAF-Net is the average AUC of M-LAF-Net. Also, the results have shown that the performance of M-DLAF-Net has been improved over the original student networks, M-DLAF-Net wo/KD. The estimated confidence maps are shown in Fig. 16.

TABLE 1
The average AUC values for MID 2006 [33], MID 2014 [34], and KITTI 2015 [35] dataset. The AUC value of ground truth confidence is measured as 'Optimal'. The results with the lowest and second lowest AUC value in each experiment are highlighted with bold and blue color.

Datasets	MID 2006 [33]		MID 2014 [34]		KITTI 2015 [35]	
	Census-SGM	MC-CNN	Census-SGM	MC-CNN	Census-SGM	MC-CNN
Haeusler et al. [21]	0.0454	0.0417	0.0841	0.0750	0.0585	0.0308
Spyropoulos et al. [12]	0.0447	0.0420	0.0839	0.0752	0.0536	0.0323
Park and Yoon [13]	0.0438	0.0426	0.0802	0.0734	0.0527	0.0303
Poggi et al. [14]	0.0439	0.0413	0.0791	0.0707	0.0461	0.0263
Kim et al. [15]	0.0430	0.0409	0.0772	0.0701	0.0430	0.0294
CCNN [18]	0.0454	0.0402	0.0769	0.0716	0.0419	0.0258
PBCP [24]	0.0462	0.0413	0.0791	0.0718	0.0439	0.0272
Shaked et al. (Conf) [6]	0.0464	0.0495	0.0806	0.0736	0.0531	0.0292
Kim et al. (conf) [17]	0.0419	0.0394	0.0749	0.0694	0.0407	0.0250
LFN [25]	0.0416	0.0393	0.0752	0.0692	0.0405	0.0253
ConfNet [26]	0.0451	0.0428	0.0783	0.0721	0.0486	0.0277
LGC-Net [26]	0.0413	0.0389	0.0735	0.0685	0.0392	0.0236
LAF-Net	**0.0405**	**0.0364**	**0.0718**	**0.0683**	**0.0385**	**0.0225**
DLAF-Net	0.0413	0.0392	0.0738	0.0698	0.0398	0.0241
Optimal	0.0340	0.0323	0.0569	0.0527	0.0348	0.0170



(a) Color images (b) Initial disparity (c) Kim et al. [17] (d) LFN [25] (e) LGC-Net [26] (f) LAF-Net (g) Ground-truth

Fig. 11. **Confidence maps on MID 2006 dataset [33] (first row) and MID 2014 dataset [34] (second row) using census-SGM and MC-CNN.** (a) color images, (b) initial disparity map, (c)-(f) are estimated confidence maps by (c) Kim et al. [17], (d) LFN [25], (e) LGC-Net [26], (f) LAF-Net, and (g) ground-truth confidence map.

TABLE 2
The average AUC values for KITTI 2015 dataset [35] with GA-Net [43]. The AUC value of ground truth confidence is measured as 'Optimal'. The results with the lowest and second lowest AUC value in each experiment are highlighted with bold and blue color.

Methods	GA-Net
Haeusler et al. [21]	0.0045
Spyropoulos et al. [12]	0.0054
Park and Yoon [13]	0.0029
Poggi et al. [14]	0.0040
CCNN [18]	0.0039
PBCP [24]	0.0041
ConfNet [26]	0.0052
LGC-Net [26]	0.0046
LAF-Net	**0.0026**
DLAF-Net	0.0033
Optimal	0.0009

TABLE 3
The average AUC values for SUN3D dataset [36] with MC-CNN [5]. The AUC value of ground truth confidence is measured as 'Optimal'. The results with the lowest and second lowest AUC value in each experiment are highlighted with bold and blue color.

Methods	MC-CNN
S-CCNN	0.1325
CCNN [18]	0.1307
CLD-MVS [28]	0.1295
LAF-Net	0.1148
M-LAF-Net	**0.1024**
M-DLAF-Net	0.1187
Optimal	0.0835

6.4 Ablation Study

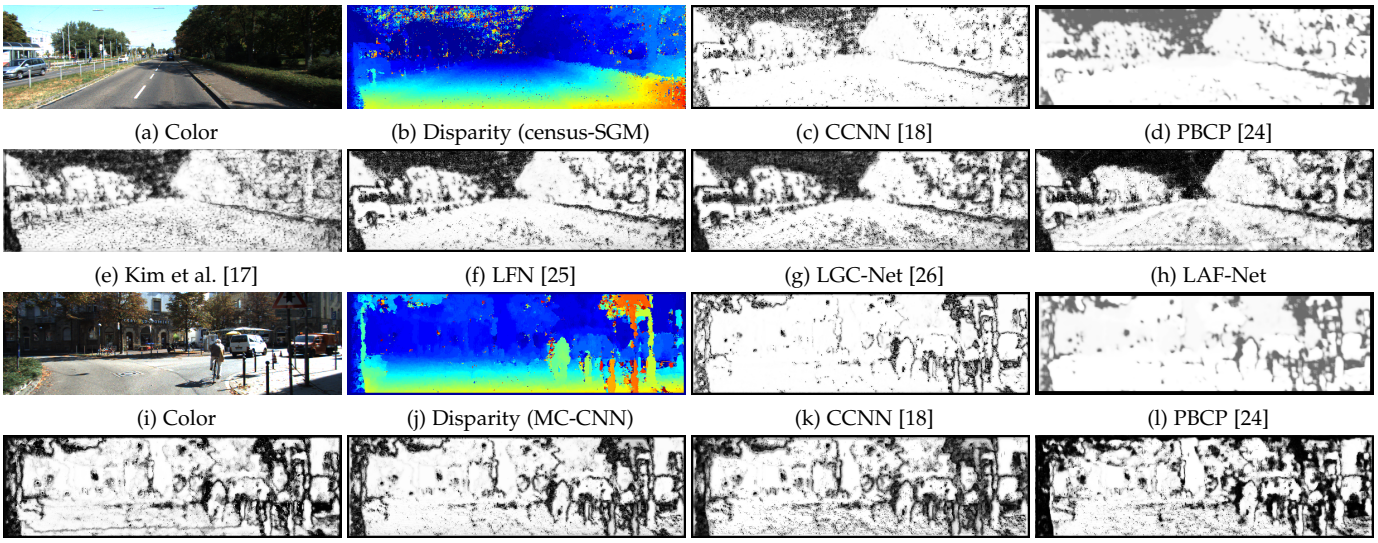
We analyzed our confidence estimation networks with the ablation evaluations, with respect to 1) various combination of different modalities such as matching cost, disparity, and color, 2) the proposed sub-networks, i.e., attention inference network, scale learning network, and recursive confidence refinement network. 3) effects of locally-varying T, and 4) effects of encoding consistency. For quantitative evaluation, we measured the average AUC values.

6.4.1 Parameter and Runtime Analysis

We first summarized the runtime and the number of parameters for the S-CCNN, CCNN [18], LAF-Net, and M-LAF-Net in Table 4. We conduct experiments on a Intel i7 CPU with NVIDIA RTX3090 GPU. Since we design S-CCNN by discarding two fully connected layers, it contains only 38K of parameters which reducing about 35 times number of parameters compared to LAF-Net and M-LAF-Net. In inference, the testing time for S-CCNN is much faster than that of LAF-Net, which demonstrate the effectiveness of the proposed framework. With compact stereo confidence estimator using knowledge distillation, we can obtain fast and accurate results which is more applicable in many real-time environments.

6.4.2 Analysis on Tri-modal Dataset

In Table 5, the ablation experiments to validate the effects of multi-modal input show the necessity of using the tri-modal input. Note that the attention inference module is not used



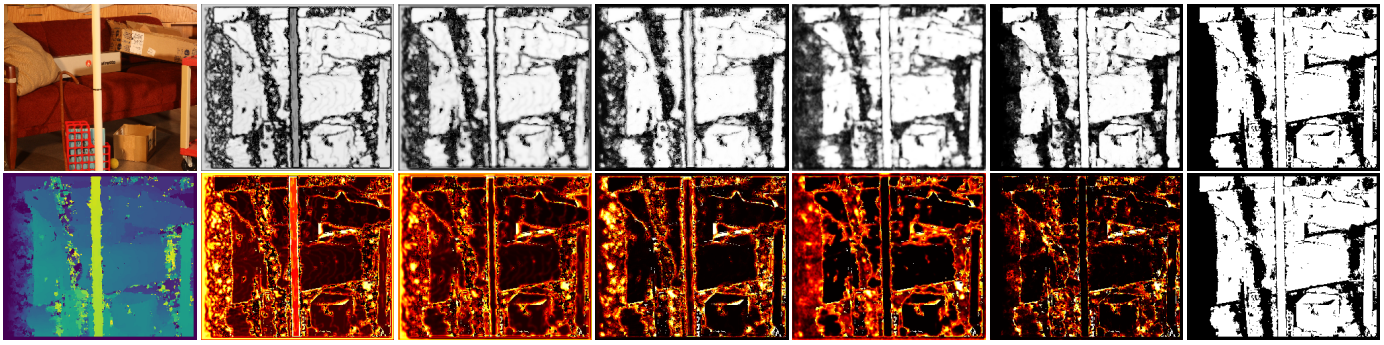
(a) Color (b) Disparity (census-SGM) (c) CCNN [18] (d) PBCP [24]

(e) Kim et al. [17] (f) LFN [25] (g) LGC-Net [26] (h) LAF-Net

(i) Color (j) Disparity (MC-CNN) (k) CCNN [18] (l) PBCP [24]

(m) Kim et al. [17] (n) LFN [25] (o) LGC-Net [26] (p) LAF-Net

Fig. 12. **Confidence maps on KITTI 2015 dataset [35] using census-SGM (first two rows), and MC-CNN (last two rows).** (From top to bottom, left to right) color images, initial disparity map, estimated confidence maps by CCNN [18], PBCP [24], Kim et al. [17], LFN [25], LGC-Net [26], and LAF-Net. In comparison to other methods, LAF-Net better estimates unreliable regions in the initial disparity maps, especially homogeneous regions.



(a) Color/Disparity (b) S-CCNN (c) CCNN [18] (d) DLAF-Net (e) DLAF-Net* (f) LAF-Net (g) Ground-truth

Fig. 13. **Confidence maps on MID 2006 dataset [62] using census-SGM:** First row shows that (a) color image and initial disparity map, estimated confidence maps by (b) S-CCNN, (c) CCNN [18], (d) DLAF-Net, (e) DLAF-Net* (DLAF-Net w/CCNN), (f) LAF-Net, and (g) ground-truth confidence map. Second row visualizes initial the difference between ground-truth and estimated confidence maps.

TABLE 4
The execution time for the DLAF-Net, CCNN, LAF-Net, and M-LAF-Net and the number of parameters of each network.

	S-CCNN	CCNN	LAF-Net	M-LAF-Net
MID 2006 (368 × 424)	2.82ms	5.65ms	42.3ms	-
MID 2014 (496 × 792)	6.78ms	15.2ms	115ms	-
KITTI 2015 (608 × 184)	2.03ms	4.35ms	38.6ms	-
SUN3D (640 × 480)	5.33ms	12.04ms	102ms	106ms
# of param.	38K	127K	1,337K	1,342K

TABLE 5
Ablation study for the various combination of input modalities in LAF-Net on MID 2006 [33], MID 2014 [34], and KITTI 2015 [35] dataset, when the raw matching cost is obtained using MC-CNN [5].

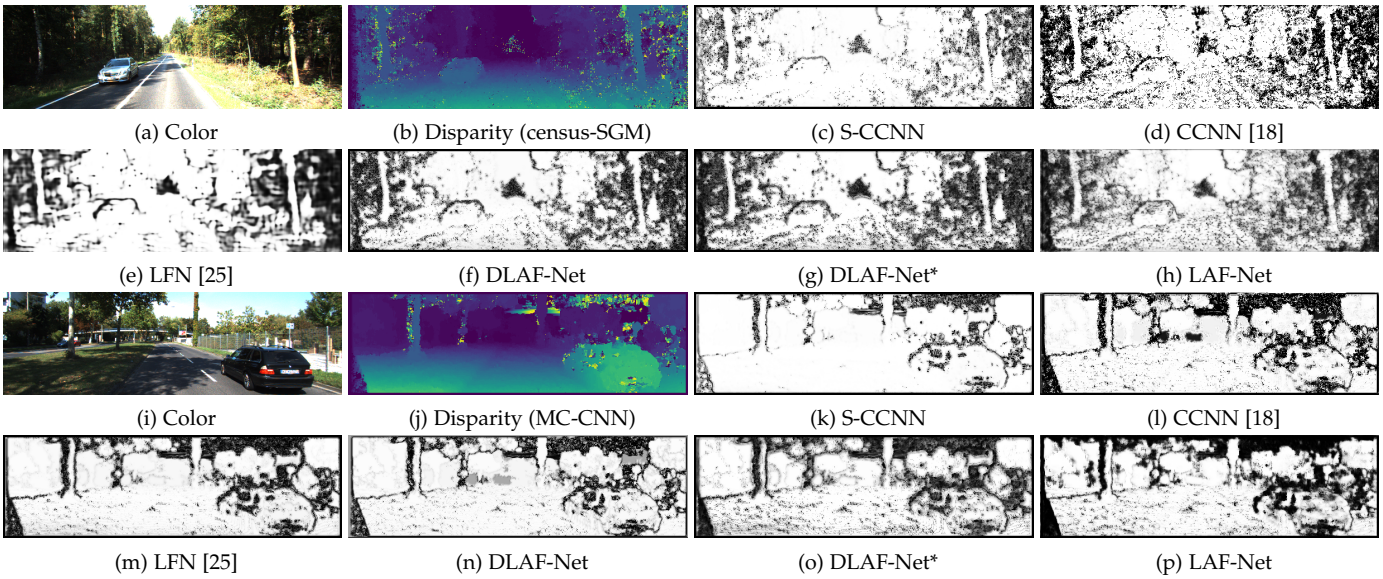
Match. cost	✓	✗	✓	✗	✓
Disparity	✗	✓	✗	✓	✓
Color	✗	✗	✓	✓	✓
MID 2006	0.0431	0.0392	0.0381	0.0375	**0.0364**
MID 2014	0.0762	0.0703	0.0687	0.0685	**0.0683**
KITTI 2015	0.0347	0.0245	0.0237	0.0231	**0.0225**

TABLE 6
Ablation study for the effectivness of each sub-network in LAF-Net on MID 2006 [33], MID 2014 [34], and KITTI 2015 [35] dataset, when the raw matching cost is obtained using MC-CNN [5]. The average AUC values for simple concatenation without fusion methods are 0.0386, 0.0689, and 0.0238 for MID 2006, MID 2014, and KITTI 2015, respectively.

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

TABLE 7
The average AUC values of DLAF-Net with MC-CNN with fixed temperature $T = 0.1, 1, 10$ and locally varying T_i.

	$T = 0.1$	$T = 1$	$T = 10$	locally varying T_i
MID 2006	0.0384	0.0385	0.0385	0.0381
MID 2014	0.0692	0.0693	0.0693	0.0690
KITTI 2015	0.0240	0.0240	0.0241	0.0237
SUN3D	0.1195	0.1196	0.1198	0.1182



(a) Color (b) Disparity (census-SGM) (c) S-CCNN (d) CCNN [18]

(e) LFN [25] (f) DLAF-Net (g) DLAF-Net* (h) LAF-Net

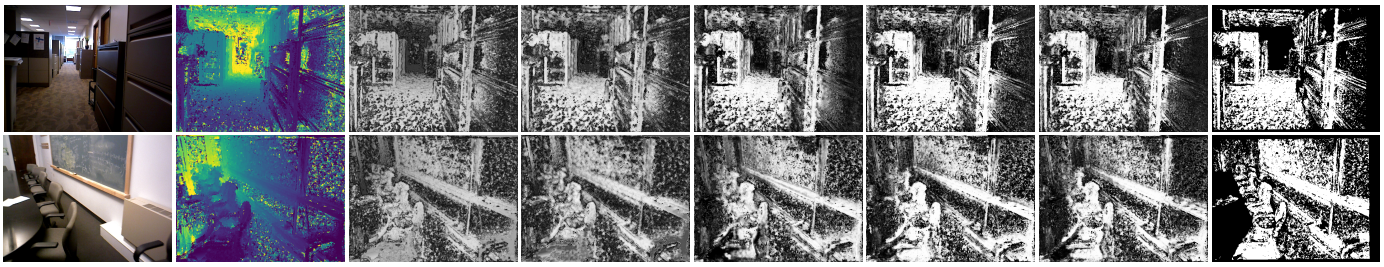
(i) Color (j) Disparity (MC-CNN) (k) S-CCNN (l) CCNN [18]

(m) LFN [25] (n) DLAF-Net (o) DLAF-Net* (p) LAF-Net

Fig. 14. **Confidence maps on KITTI 2015 dataset [35] using census-SGM (first two rows), and MC-CNN (last two rows).** (From top to bottom, left to right) color image, initial disparity map, estimated confidence maps by S-CCNN, CCNN [18], LFN [25], DLAF-Net, DLAF-Net* (DLAF-Net w/CCNN), and LAF-Net.

(a) Color image (b) Initial disparity (c) LGC-Net [26] (d) DLAF-Net w/CCNN (e) LAF-Net

Fig. 15. **Confidence maps on KITTI 2015 dataset [35] using GA-Net:** (a) color image, (b) initial disparity map, estimated confidence maps by (c) LGC-Net [26], (d) DLAF-Net w/CCNN, and (e) LAF-Net.



(a) Color (b) Depth map (c) S-CCNN (d) CCNN [18] (e) M-DLAF-Net (f) LAF-Net (g) M-LAF-Net (h) Ground-truth

Fig. 16. **Confidence maps on SUN3D dataset [36] using MC-CNN.** (a) Color image, (b) estimated depth map with MC-CNN, (c)-(g) are estimated confidence maps by (c) S-CCNN, (d) CCNN [18], (e) M-DLAF-Net w/S-CCNN (1st rows) and w/CCNN (2nd rows), (f) LAF-Net, (g) M-LAF-Net, and (h) ground-truth confidence map.

TABLE 8
Ablation study for the geometric and photometric consistency in M-LAF-Net on SUN3D dataset [36].

Geometric cons.	✗	✓	✗	✓
Photometric cons.	✗	✗	✓	✓
SUN3D	0.1148	0.1057	0.1098	0.1024

for input of single modality. Although the bi-modal input improved the ability to predict reliable pixels, the full usage of tri-modal input shows the best performance.

6.4.3 The Effects on Fusion Methods

In Table 6, ablation experiments to validate the effects of the proposed fusion methods. Compared to the simple concatenation technique, the confidence estimator is improved with the attention and scale obtained from the attention and scale inference networks. Also, the recursive confidence refinement networks show the additional improvement.

6.4.4 The Effects on Locally Varying T

We analyzed the effectiveness of our temperature inference networks with the ablation evaluations. The average AUC values of confidence maps by estimating DLAF-Net using MC-CNN are shown in Table 7. With locally varying T, we enhanced confidence maps by transferring more informative value to student networks.

6.4.5 The Effects on Encoding Consistency

In Table 8, we performed ablation experiments to validate the effects of the proposed geometric and photometric consistency in M-LAF-Net. By leveraging the information of the consistency of multiple inputs, the ability to predict unreliable pixels can be effectively improved.

7 Conclusion

We presented LAF-Net that estimates confidence with tri-modal input, including matching cost, disparity, and color image through deep networks. The key idea of the proposed

method is to design locally adaptive attention and scale inference networks to generate optimal fusion weights. In addition, the confidence estimation performance is further improved with recursive refinement networks. In addition, we presented an effective confidence estimator through knowledge distillation using the LAF-Net taking tri-modal input as teacher, where we learned the locally varying temperature which is effective in transferring more informative value to student networks. The proposed method has competitive accuracy with simpler networks than teacher. We further applied the proposed framework to multiview stereo confidence estimation which demonstrates the generalization ability of the proposed framework. A direction for further study is to examine how confidence estimation networks could be learned in an unsupervised manner.

Acknowledgments This research was supported by the Yonsei University Research Fund of 2021 (2021-22-0001).

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, 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, 2008.

[3] S. Kim, B. Ham, B. Kim, and K. Sohn, "Mahalanobis distance cross-correlation for illumination invariant stereo matching," vol. 24, no. 11, pp. 1844–1859, 2014.

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

[5] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1592–1599, Jun. 2015.

[6] A. Shaked and L. Wolf, "Improved stereo matching with constant highway networks and reflective confidence learning," *IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2017.

[7] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang, "On building an accurate stereo matching system on graphics hardware," pp. 467–474, Nov. 2011.

[8] 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.

[9] 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, 2002.

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

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

[12] A. Spyropoulos, N. Komodakis, and P. Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1621–1628, Jun. 2014.

[13] M. Park and K. Yoon, "Leveraging stereo matching with learning-based confidence measures," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 101–109, Jun. 2015.

[14] 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," *3*.

[15] 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, 2017.

[16] 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.

[17] 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, 2019.

[18] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," *Brit. Mach. Vis. Conf.*, vol. 10, Sep. 2016.

[19] 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, 2012.

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

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

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

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

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

[25] Z. Fu and M. A. Fard, "Learning confidence measures by multi-modal convolutional neural networks," pp. 1321–1330, 2018.

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

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

[28] Z. Li, W. Zuo, Z. Wang, and L. Zhang, "Confidence-based large-scale dense multi-view stereo," *IEEE Trans. Image Process.*, vol. 29, pp. 7176–7191, 2020.

[29] S. Im, H. G. Jeon, S. Lin, and I. S. Kweon, "Dpsnet: End-to-end deep plane sweep stereo," *Int. Conf. Learn. Represent.*, 2019.

[30] A. Kuhn, C. Sormann, M. Rossi, O. Erdler, and F. Fraundorfer, "Deepc-mvs: Deep confidence prediction for multi-view stereo construction," 2019.

[31] J. Zhang, Y.Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," *Brit. Mach. Vis. Conf.*, 2020.

[32] J. Zhang, Y.Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Learning stereo matchability in disparity regression networks," *Int. Conf. Pattern Recog.*, 2020.

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

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

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

[36] J. Xiao, A. Owens, and A. Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," *Int. Conf. Comput. Vis.*, pp. 1625–1632, 2013.

[37] S. Kim, S. Kim, D. Min, and K. Sohn, "Laf-net: Locally adaptive fusion networks for stereo confidence estimation," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 205–214, 2019.

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

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

[40] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Deep Learning and Representation Learning Workshop: NIPS*, 2014.

[41] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9163–9171, 2019.

[42] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," *Eur. Conf. Comput. Vis.*, 2018.

[43] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 185–194, 2019.

[44] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *Int. Conf. Learn. Represent.*, 2015.

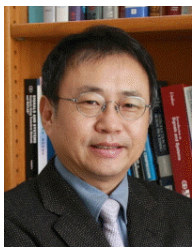
[45] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *Int. Conf. Learn. Represent.*, 2017.
[46] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5–6, pp. 544–557, 2009.
[47] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2023–2049, 2015.
[48] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 826–834, 2016.
[49] D. Lopez-Paz, L. Bottou, B. Scholkopf, and V. Vapnik, "Unifying distillation and privileged information," *Int. Conf. Learn. Represent.*, 2016.
[50] Z. Luo, J. T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with priviledged modalities," *Eur. Conf. Comput. Vis.*, 2018.
[51] X. Jia, B. D. Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," *Adv. Neural Inform. Process. Syst.*, pp. 667–675, Dec. 2016.
[52] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Adv. Neural Inform. Process. Syst.*, pp. 2017–2025, Dec. 2015.
[53] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," *Adv. Neural Inform. Process. Syst.*, pp. 2414–2422, Dec. 2016.
[54] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[55] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5483–5492, 2019.
[56] T. Khot, S. Agrawal, S. Tulsiani, C. Mertz, S. Lucey, and M. Hebert, "Learning unsupervised multi-view stereopsis via robust photometric consistency," *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
[57] Q. Xu and W. Tao, "Learning inverse depth regression for multi-view stereo with correlation cost volume," *AAAI*, 2020.
[58] ——, "Pvsnet: Pixelwise visibility-aware multi-view stereo network," *arXiv*, 2020.
[59] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.
[60] M.Poggi, S. Kim, F. Tosi, S. Kim, F. Aleotti, D. Min, K. Sohn, and S. Mattoccia, "On the confidence of stereo matching in a deep-learning era: a quantitative evaluation." *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[61] 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.*, pp. 611–625, Oct. 2012.
[62] [Online] http://vision.middlebury.edu/stereo/.

Sunok Kim received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2014 and 2019. From 2019 to 2021, she has been Post-Doctoral Researcher in School of Electrical and Electronic Engineering at Yonsei University. Since 2021, she has been an assistant professor with the Department of Software, Korea Aerospace University, Goyang. Her current research interests include 2D/3D computer vision, machine learning, and artificial intelligence.

Seungryong Kim received the B.S. and Ph.D. degrees from the School of Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2012 and 2018, respectively. From 2018 to 2019, he was Post-Doctoral Researcher in Yonsei University, Seoul, Korea. From 2019 to 2020, he has been Post-Doctoral Researcher in School of Computer and Communication Sciences at 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 2D/3D computer vision, computational photography, and machine learning.

Dongbo Min received the BS, MS, and PhD 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, Massachusetts, US. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. From 2015 to 2018, he was an assistant professor in the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been in the Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea. His current research interests include computer vision, deep learning, and video processing.

Pascal Frossard (Fellow, IEEE) was a member of the Research Staff with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, from 2001 to 2003. He has been a Faculty with the École PolytechniqueFéd érale de Lausanne (EPFL) since 2003, where he currently heads the Signal Processing Laboratory (LTS4). His research interests include network data analysis, image representation and understanding, and machine learning. He is a fellow of ELLIS. He was a recipient of the Swiss NSF Professorship Award in 2003, the IBM Faculty Award in 2005, the IBM Exploratory Stream Analytics Innovation Award in 2008, the Google Faculty Award in 2017, the IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award in 2011, and the IEEE Signal Processing Magazine Best Paper Award in 2016.



Kwanghoon Sohn received the B.E. degree in electronic engineering from Yonsei University, Seoul, 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, 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.