IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2019, accepted October 22, 2019, date of publication November 4, 2019,
date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950924

Learning Efficient Stereo Matching Network With Depth Discontinuity Aware Super-Resolution

CHENGGANG GUO, DONGYI CHEN, AND ZHIQI HUANG
School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Dongyi Chen (dychen@uestc.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002803, and in part by the National Natural Science Foundation of China under Grant 61572110.

ABSTRACT Deep convolutional neural networks (CNNs) have shown great potential to provide accurate depth estimation based on stereo images. Previous work has focused on developing robust stereo matching architectures, while little attention has been paid on improving the network efficiency. In this paper, we propose an efficient Siamese CNN architecture that combines the low resolution disparity estimation and the depth discontinuity aware super-resolution. Specifically, we propose to construct, filter and perform regression on a low resolution cost volume through the designed stereo matching backbone network. A fast depth discontinuity aware super-resolution subnetwork is proposed for upsampling the low resolution disparity map to the desired resolution. Under the guidance of the intensity edge features extracted from the left color image, depth edge residuals are hierarchically learned to refine the upsampled depth map. A delayed upsampling structure is designed to ensure that the computational complexity is proportional to the spatial size of the input disparity map. We also propose to supervise the first derivative loss of the predicted disparity map that makes the network adaptively aware of the depth discontinuity edges. Experiments show that the proposed stereo matching network achieves a comparable prediction accuracy and much faster running speed compared with state-of-the-art methods.

INDEX TERMS Stereo matching network, disparity estimation, depth map super-resolution, depth discontinuity aware loss.

I. INTRODUCTION

Depth estimated from stereo images has been the core information for vision-based practical applications, such as obstacle avoidance for robot navigation [1], 3D scene reconstruction for augmented and virtual reality system [2], and 3D visual object tracking and location [3], [4]. Given a pair of pre-rectified stereo images, the target of stereo matching is to accurately compute a disparity value for each pixel in the reference image. According to the taxonomy concluded by Scharstein et al. [5], traditional stereo matching algorithms typically include four consecutively performed steps: matching cost computation, cost aggregation, disparity computation and disparity refinement.

In recent years, with the rapid development of deep learning, lots of convolutional neural network (CNN) based methods have been proposed to solve the stereo matching problem, since the milestone work of MC-CNN [6]. Early deep stereo networks are designed to learn similarity metrics from a large set of cropped image patches [6]–[10]. Regularization or global optimization approaches, such as semi-global matching (SGM) [11], left-right consistency checks and Markov Random Field (MRF) [10], are formulated as post-processing models. Later, many deep stereo networks attempt to directly learn various stereo matching regression functions end-to-end without the need of adding post-processing. In GC-Net [12], a fully differentiable 4D cost volume (feature channel $(C)\times$max disparity $(\frac{1}{2}D)\times$feature height $(\frac{1}{2}H)\times$feature width $(\frac{1}{2}W)$) is formed for the first time. A 3D convolutional architecture is utilized to filter and refine this 4D representation. Following the pipeline of GC-Net, PSMNet [13] and GwcNet [14] exploit multiscale context aggregation on their 4D cost volumes $(C \times \frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W)$ by stacking hourglass modules, which is equivalent to stacking 3D light-weighted encoder-decoder structures.

Although applying 3D convolutions on the 4D cost volume can better aggregate neighboring disparities and produce

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

lower error rates, the added computational burden and memory footprint of the additional dimension makes training and prediction relatively slow, especially for high resolution stereo image pairs. Currently, there are two ideas that try to tackle this issue. The first one is to replace the 3D convolution operator with a differentiable approximation of the classical optimization method, like the semi-global aggregation layer proposed in [15]. The second one is to construct a low resolution cost volume with a large downsampling factor (e.g. $C \times \frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W$ or $C \times \frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W$) and apply 3D convolutions on this cost volume, like the deep stereo networks proposed in [16]. Consistent with the second idea, we also believe that the operation resolution of 3D convolutions is the key factor that affects the efficiency for predicting high resolution disparity maps.

In this paper, we propose an efficient end-to-end stereo matching network that predicts high resolution depth maps with state-of-the-art accuracy and comparable running speed. Overall, our proposed deep stereo network filters and regresses a low resolution cost volume, and hierarchically upsamples the initial low resolution disparity map to a higher resolution under the guidance of the left color image. We draw on the recent research work of depth map super-resolution and propose a fast depth discontinuity aware super-resolution subnetwork for upsampling the predicted low resolution disparity map to the desired resolution. The super-resolution subnetwork serves as a high frequency hierarchical upsampling and refinement module that performs end-to-end joint training with the stereo matching backbone network. We aim to build a super-resolution subnetwork that reveals the connections between the intensity edges of high resolution color image and the depth discontinuity edges of the disparity map. Specifically, we extract high frequency information from each channel of the left color image, and then feed it to a set of downsampling convolutional layers to obtain a guidance pyramid. For every upsampling level, the guidance of corresponding level from the guidance pyramid is fused with the high frequency information extracted from the input disparity map. The high frequency refinement residuals are learned at the resolution level of the input disparity map until the deconvolution layer placed at the end of the subnetwork completes the residual upsampling, which provides a learnable upsampling kernel compared to the direct disparity interpolation. The learned high frequency residuals are added back to the input disparity map to refine the high frequency details missed in the upsampling process. By doing the residual learning at the resolution level of the input disparity map, the computational complexity is only proportional to the spatial size of the input disparity map. Additionally, we also propose a depth discontinuity aware loss that gives supervision on the first derivatives of the predicted disparity maps. Experiments verify that the accuracy of existing end-to-end stereo matching networks can be further improved using the proposed depth discontinuity aware loss without modifying architectures.

Our main contributions can be summarized as below:

Firstly, we propose a fast depth discontinuity aware super-resolution subnetwork for low resolution disparity map upsampling and refinement. The proposed subnetwork and the stereo matching backbone network are jointly trained end-to-end to enable an efficient and accurate prediction of high resolution disparity map.

Secondly, we propose to supervise the loss between the first derivatives of the estimated disparity maps and the first derivatives of the groundtruth disparity maps. The proposed depth discontinuity aware loss provides effective supervision of depth discontinuity edges and only requires groundtruth disparity maps.

Thirdly, experimental results on several large scale benchmarks show that the proposed stereo matching network achieves a comparable prediction accuracy and much faster running speed compared with the state-of-the-art deep stereo matching methods.

II. RELATED WORK

This section begins with a brief review of a number of end-to-end stereo matching networks in recent years. The key innovations proposed inside these deep stereo models are identified and discussed. Since our proposed efficient deep stereo model utilizes a super-resolution subnetwork to restore the predicted low resolution disparity map into a refined high resolution disparity map, some recent depth map super-resolution methods which train end-to-end deep networks are also briefly introduced. The innovations of our proposed deep stereo model are discussed against the previous works in the final subsection.

A. END-TO-END STEREO MATCHING NETWORKS

End-to-end deep stereo networks have not been extensively studied until the first large scale synthetic stereo datasets disclosed by Mayer *et al.* [17]. In addition to the release of the datasets, they also proposed a DispNetC architecture for disparity estimation, which has a contractive part and an expanding part with long-range links. DispNetC explicitly uses a 1D correlation layer that horizontally correlates left and right features. Using an improved architecture of DispNetC as the first stage, Pang *et al.* [18] introduced a multiscale residual learning scheme (CRL) for the second stage refinement. Kendall *et al.* [12] proposed to construct a 4D cost volume for context aggregation for the first time. In their pioneering work (GC-Net), 3D convolutions are employed to filter the cost volume over height $\times$ width $\times$ disparity dimensions. They also proposed a differentiable soft argmin operation to directly regress subpixel disparities from the cost volume. Chang and Chen [13] proposed the pyramid stereo matching network (PSMNet), which explores the context aggregation of the feature extraction stage and cost volume filtering stage by using a pyramid pooling module and a stacked hourglass 3D CNN, respectively. Guo *et al.* [14] focused on the formation of cost volume and considered improving the quality of cost volume by designing a robust correlation metric.

They proposed to construct the cost volume by group-wise correlations (GwcNet). Then the hybrid cost volume, built by combining feature concatenation volume and group correlation volume, is fed into an improved version of stacked 3D hourglass modules. Yu *et al.* [19] proposed a learning-based cost aggregation sub-architecture that select the most possible aggregated cost proposals. Zhang *et al.* [15] proposed two novel substitution layers (GA-Net) for the computation of the costly 3D convolutional layer. The first is formulated by a differentiable approximation of the semi-global matching, and the second, called the local guided aggregation layer, functions substantially similar to the 3D convolutional layer except that it separates the aggregation calculations in the spatial and the depth dimensions. Tulyakov *et al.* [20] proposed a novel sub-pixel cross-entropy loss with a MAP estimator to make their network applicable to different disparity ranges without re-training (PDS). Poggi *et al.* [21] proposed to use a small amount of additional sparse depth measurements to improve the domain shift ability of pre-trained stereo networks.

Some work has tried to learn end-to-end stereo matching networks by designing various multitask structures. Du *et al.* [22] proposed to learn the foreground-background segmentation as an auxiliary task to reinforce the disparity estimation (FBA-AMNet). Song *et al.* [23] proposed to integrate edge information into the disparity learning process by utilizing an edge sub-network (EdgeStereo). An edge-aware smoothness loss is defined to encourage disparities to be locally smooth and penalizes depth changes in non-edge regions. Yang *et al.* [24] employed semantic features from segmentation and introduced a semantic softmax loss for disparity estimation (SegStereo).

While most existing deep stereo models put their emphasis on improving the prediction accuracy, a few efforts focus on improving the network matching speed for practical applications. Khamis *et al.* [16] presented a real-time stereo matching network (StereoNet) that forms a low resolution cost volume to reduce the computation of 3D convolutions. Their core ideas are that the cost volume can be filtered at a coarse resolution, and the regressed low resolution disparity map can then be hierarchically upsampled to a high resolution disparity map. Tonioni *et al.* [25] developed a novel real time self-adaptive network (MADNet) that address the domain shift issue by independently training sub-portions of the network. Chabra *et al.* [26] followed the work of StereoNet, and proposed a novel disparity refinement network that takes geometric error, photometric error and unrefined disparity as input and produces the refined disparity map (StereoDRNet).

B. END-TO-END DEPTH MAP SUPER-RESOLUTION NETWORKS

For the problem of depth sensing, it is desirable to upsample the low resolution depth map provided from a low-cost depth camera by employing super-resolution techniques. Recently deep learning-based super-resolution methods like the SRCNN [27] and FSRCNN [28] proposed by Dong et al. have drawn much attention due to superior performance. But unlike the general image super-resolution problem, the depth map super-resolution has its own characteristics, e.g. less texture and sharp boundary. Typically, methods of depth map super-resolution can be divided into two categories: the methods with only depth maps as input [29], and the methods using low resolution depth maps and high resolution color images [30]–[35]. Since the high resolution color image aligned to the depth map is available in the setting of stereo matching, we only review a few end-to-end depth map super-resolution networks that takes a high resolution color image as a guidance.

Hui *et al.* [30] proposed a multi-scale guided network (MSG-Net) for depth map super resolution that complements low resolution depth features with high resolution intensity features using a multi-scale fusion strategy. Zhao *et al.* [31] designed a conditional generative network (CDcGAN) for resolving the problem of simultaneous color image and depth image super-resolution. Ni *et al.* [32] proposed a dual-stream CNN that integrated the color and depth information. Edge map generated by the high resolution color image and low resolution depth map is taken as additional information for disparity refinement. Song *et al.* [33] exploited the depth field statistics and the local correlation between depth image and color image for the further refinement of the learned depth image. Zhou *et al.* [34] observed that color images help to learn the super-resolution network of depth maps contaminated by noises. Wen *et al.* [35] proposed a coarse-to-fine CNN architecture and presented a data-driven filter method to approximate the ideal filter for depth super-resolution.

C. INNOVATIONS COMPARING WITH PREVIOUS DEEP STEREO MODELS

Our work follows state-of-the-art stereo matching deep networks such as GC-Net [12], PSMNet [13], and GwcNet [14], all of which build robust 4D cost volumes and apply 3D convolutions for context aggregation. However, the high computation burden brought by the 3D convolution operations makes it hard for existing networks to find a good balance between inference speed and prediction accuracy. In this paper, we propose an end-to-end CNN architecture that combines the low resolution disparity estimation and the depth discontinuity aware super-resolution. The initial noise disparity map is regressed from the low resolution cost volume, which only consumes a low computation cost. Then the initial low resolution disparity map is gradually upsampled to a high resolution under the guidance of corresponding left color image. Comparing to the deep stereo models exploiting auxiliary task, such as FBA-AMNet [22] and EdgeStereo [23], we propose to learn a subnetwork of high frequency features from the high resolution color image that provide necessary high frequency details for the super-resolution process. We regard the super-resolution of low resolution depth map as a high frequency cascade refinement process, which benefits from the concept of spectral decomposition from MSG-Net [30]. At every scale, the high frequency residuals

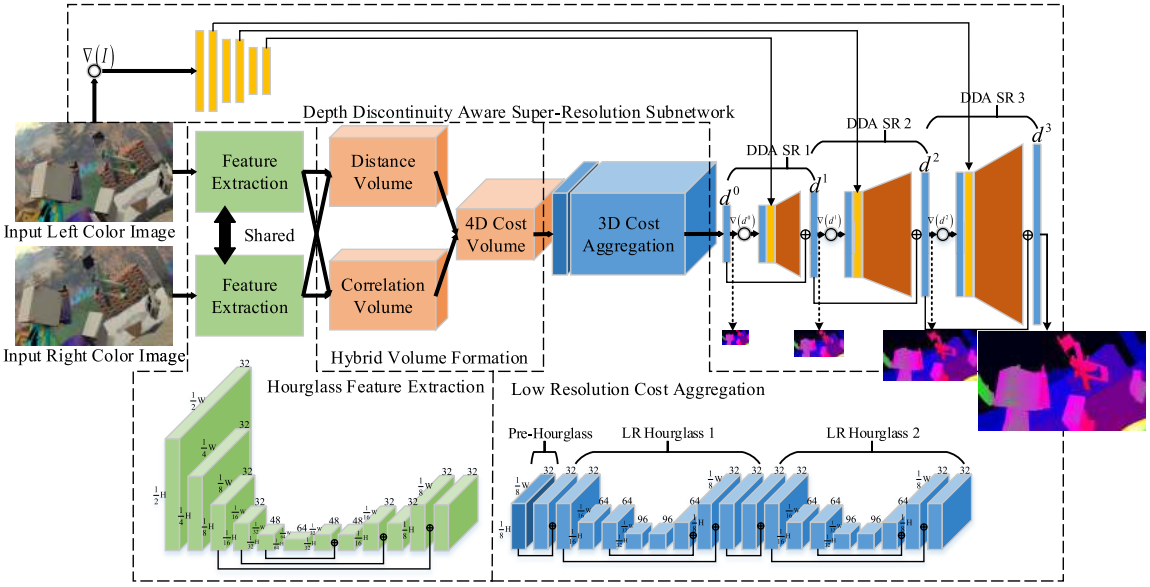

FIGURE 1. The architecture of our efficient stereo matching network (ESMNet). Each subnetwork is denoted by a different color and separated by the black dashed line. In the hourglass feature extraction subnetwork and the low resolution cost aggregation subnetwork, we have annotated the resolution level and the number of feature channels of each convolution layer. The black circle denotes the operation of extracting high frequency information for the color image and the disparity maps.

are learned in the super-resolution subnetwork and then added back to the upsampled depth map to rectify the incorrect high frequency component. This structure of explicit learning high frequency residuals differs from the DispResNet in CRL [18], the Residual Pyramid in EdgeStereo [23] and the refinement structure in StereoNet [16] and StereoDRNet [26]. Finally, different from the l_1 loss adopted in [12], [18], [19], [24] and the smooth l_1 loss adopted in [13]–[15], [22], the end-to-end learning of our deep stereo model is supervised by our proposed depth discontinuity aware loss, which gives extra supervision with first order derivatives about the output depth maps. The difference between our super-resolution subnetwork and the existing depth map super-resolution models is that not only should the depth map be upsampled to recover the details of the depth discontinuous edges, but also the smooth regions of the depth continuity should be refined.

III. EFFICIENT STEREO MATCHING NETWORK

In this section, we first introduce the architecture of our efficient stereo matching network, and named ESMNet for convenience. Then the proposed depth discontinuity aware super-resolution subnetwork is described in detail. Finally, the proposed first order derivative loss that makes the network adaptively aware the depth discontinuity edges is discussed.

A. NETWORK ARCHITECTURE

Fig. 1 illustrates the architecture of our ESMNet. Each subnetwork that functions similarly to the concluded steps in [5] is separated by the black dashed line. The stereo matching backbone network adopt the designing principles from StereoNet [16] that mainly includes a low resolution cost volume aggregation module and a hierarchical disparity upsampling module. However, we make several key modifications that help to improve the stereo matching accuracy with no decrease of running speed. The details of the proposed stereo matching backbone architecture are listed in the Table 1.

For the feature extraction Siamese subnetwork, we use a hourglass shaped architecture to encode global context information. As a strong competitor for hourglass architectures, atrous convolution has been widely applied to encode multiscale contextual information for stereo matching, e.g. the AM module proposed in AMNet [22]. But we aim to design a computationally efficient feature extractor by reducing the feature resolution. The input color image pair can be quickly reduced to a very low level of operating resolution by continuous downsampling convolutions. As the resolution decreases, we increase the feature channels moderately for each resolution level to encode more features. Long range skip connections are added between the encoder part and the decoder part to enable information complement. The deconvolution operator is utilized to restore the feature resolution to the desired resolution level. For the first three successive downsampling convolution layers, the kernel size is set to 5×5 to obtain a relative rich local patterns. For the rest three downsampling convolution layers, the kernel size is set to 3×3. In total, we have six downsampling layers. For every downsampling convolution layer, the stride is set to 2. Batch normalization and ReLU activations are applied. A 3×3 convolution layer without batch normalization or activation is applied as the output layer of the hourglass feature extraction subnetwork. Finally, this feature extraction subnetwork outputs intermediate semantic features of size $C \times \frac{1}{8}H \times \frac{1}{8}W$. Thus, the total downsampling factor is 2^3.

For the cost volume formation, we construct a hybrid metric volume that takes in the coarse stereo correspondences calculated by the group-wise correlation metric

TABLE 1. Structure details of the proposed stereo matching backbone architecture. *H*, *W* represents the height and the width of the input image. BN denotes the batch normalization. C denotes the number of output channels for the convolution layer. S1/2 denotes the convolution stride. DConv denotes the deconvolution operation.

Index	Layer Description	Output Size
(1)	RGB Image	3xHxW
	Hourglass Feature Extraction	
(2)	[Conv5x5, C32, S2], BN, ReLU	32xH/2xW/2
(3)	[Conv5x5, C32, S2], BN, ReLU	32xH/4xW/4
(4)	[Conv5x5, C32, S2], BN, ReLU	32xH/8xW/8
(5)	[Conv3x3, C32, S2], BN, ReLU	32xH/16xW/16
(6)	[Conv3x3, C48, S2], BN, ReLU	48xH/32xW/32
(7)	[Conv3x3, C64, S2], BN, ReLU	64xH/64xW/64
(8)	ReLU{[DConv3x3, C48, S2], BN+(6)}	48xH/32xW/32
(9)	[Conv3x3, C48, S1], BN, ReLU	48xH/32xW/32
(10)	ReLU{[DConv3x3, C32, S2], BN+(5)}	32xH/16xW/16
(11)	[Conv3x3, C32, S1], BN, ReLU	32xH/16xW/16
(12)	ReLU{[DConv3x3, C32, S2], BN+(4)}	32xH/8xW/8
(13)	[Conv3x3, C32, S1]	32xH/8xW/8
	Hybrid Metric Cost Volume	
(14)	Correlation Sub-Volume	16xD/8xH/8xW/8
(15)	Distance Sub-Volume	32xD/8xH/8xW/8
(16)	From (14),(15): Concat	48xD/8xH/8xW/8
	Low Resolution (LR) Cost Aggregation	
(17)	[Conv3x3x3, C32, S1], BN, ReLU	32xD/8xH/8xW/8
	LR Hourglass 1,2	
(18)	ReLU{[Conv3x3x3,C32,S1],BN+Input}	32xD/8xH/8xW/8
(19)	[Conv3x3x3, C64, S2], BN, ReLU	64xD/16xH/16xW/16
(20)	[Conv3x3x3, C64, S1], BN, ReLU	64xD/16xH/16xW/16
(21)	[Conv3x3x3, C96, S2], BN, ReLU	96xD/32xH/32xW/32
(22)	[Conv3x3x3, C96, S1], BN, ReLU	96xD/32xH/32xW/32
(23)	[DConv3x3x3,C64, S2], BN	64xD/16xH/16xW/16
(24)	From (20): [Conv1x1x1,C64,S1], BN	64xD/16xH/16xW/16
(25)	ReLU{(23)+(24)}	64xD/16xH/16xW/16
(26)	[DConv3x3x3,C32,S2], BN	32xD/8xH/8xW/8
(27)	From (18): [Conv1x1x1,C32,S1], BN	32xD/8xH/8xW/8
(28)	ReLU{(26)+(27)}	32xD/8xH/8xW/8
	Initial LR Disparity Output	
(29)	[Conv3x3x3, C1, S1]	1xD/8xH/8xW/8
(30)	Disparity Regression: d^0	1xH/8xW/8

(proposed in [14]) and the absolute distance metric. The hybrid metric volume can provide an good initial guess for stereo correspondences and makes the training converge faster than the classical left-right feature concatenation based cost volume [12], [13], [19]. Let C denotes the number of output feature channels in the feature extraction subnetwork. The left features F_l and right features F_r are evenly divided into G groups along the feature channel dimension. The correlation sub-volume is computed at every group g and disparity level d, which is

$$C_{corr}(g, d, h, w) = \frac{1}{C/G} \left\langle F_l^g(h, w), F_r^g(h, w-d) \right\rangle \quad (1)$$

where $\langle \cdot \rangle$ denotes the inner product. The absolute distance metric is computed at every feature channel c and disparity level d, which is

$$C_{dis}(c, d, h, w) = -\left| F_l^c(h, w) - F_r^c(h, w-d) \right| \quad (2)$$

Finally, the two metric sub-volumes are stacked along the channel dimension to form the hybrid metric volume, which has a size of $(C+G) \times \frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W$.

Typically, it is necessary to filter and refine the constructed coarse cost volume by applying the convolution operation along the height, width and disparity dimensions. After a series of filtering and aggregation operations, the coarse cost volume that contains non-smooth and incorrect disparity noises can be refined. We find that 19, 24 and 30 3D convolutional operations are used for the cost volume aggregation in GC-Net [12], PSMNet [13] and GwcNet [14], respectively. Moreover, some of the 3D convolutional operations are conducted at a relative higher resolution ($\geq \frac{1}{4}H \times \frac{1}{4}W$), which occupy a considerable amount of computation time. In contrast, we apply a limit number of 3D convolutional operations at a very low operation resolution to reduce the time of cost aggregation. As shown in Fig. 1, we make some modifications to the hourglass structure proposed by GwcNet. For the pre-hourglass convolution, we only use one layer of 3D convolution to fuse and transform metric representations. Then two 3D hourglass modules are stacked, each of which contains 9 3D convolutions. For each hourglass module, the input cost volume is first filtered at the resolution of $\frac{1}{8}W \times \frac{1}{8}H$ by a single 3D convolution layer. Then it is continuous downscaled to the resolution of $\frac{1}{32}W \times \frac{1}{32}H$ by four 3D convolutions. Two 3D deconvolution layers are responsible for upsampling the downscaled cost volume. One $1 \times 1 \times 1$ 3D convolution layer is utilized as a skip connection after each deconvolution operation. Overall, our design maintains the operation resolution of 3D convolution at a very low level, which consumes less computation time.

The initial low resolution disparity map is regressed from the filtered cost volume by using the disparity regression operation [12]:

$$d^0 = \sum_{d=0}^{d_{max}} d \times \sigma(C^A(d)) \quad (3)$$

where $\sigma(\cdot)$ is the softmax operation, and C^A is the cost volume filtered by the low resolution cost aggregation module.

B. DEPTH DISCONTINUITY AWARE SUPER-RESOLUTION SUBNETWORK

Fig. 2 gives a more detailed illustration of the proposed fast super-resolution subnetwork. The structure details are listed in Table 2. As discussed in the related work section, we adopt the spectral decomposition idea from the work of depth map super-resolution [30]. For depth maps, the high frequency information corresponds to the depth discontinuity edges and the low frequency information corresponds to the depth flat regions. Different upsampling strategies for different frequency components are adopted in the proposed super-resolution subnetwork. Recent works have pointed out that edge cues from the color image have semantic connections with depth discontinuity edges from the depth map. Therefore, we perform an early frequency spectrum decomposition from both of the left color image and the depth maps. The high frequency components extracted from both of the color image and the depth map are feed into a subnetwork for learning depth discontinuity residuals. The learned residuals are then added back to the bilinear upsampled depth map for high frequency refinement. From the perspective of network structure, the proposed super-resolution subnetwork

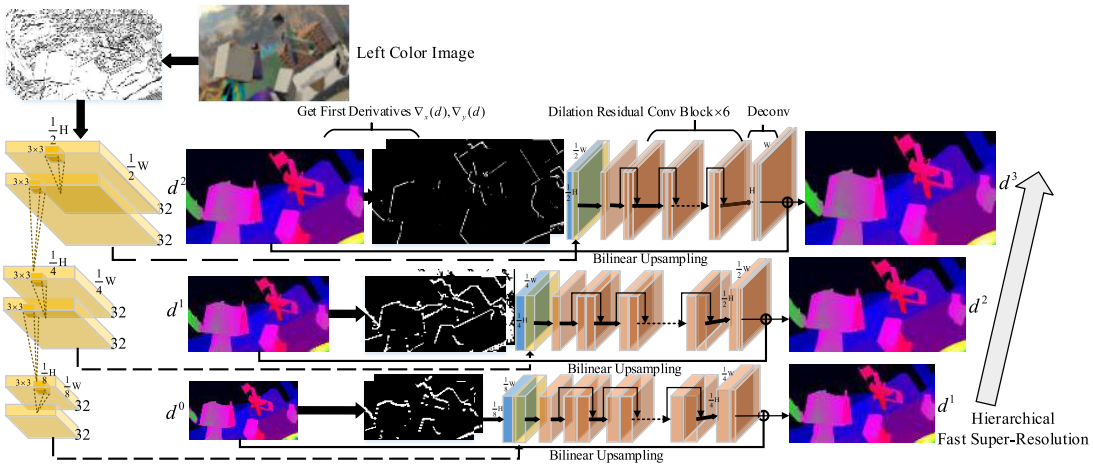

FIGURE 2. The depth discontinuity aware super-resolution subnetwork. Depth maps are colored for better visualization. The first derivatives of each depth map in the x and y directions are extracted by applying the Sobel kernel.

TABLE 2. Structure details of the proposed depth discontinuity aware super-resolution subnetwork. Dil denotes the convolutional dilation rate.

Index	Layer Description	Output Size
	Intensity Edges Extraction Module	
(1)	From Left Image: Conv-Sobel, BN	6xHxW
(2)	[Conv3x3, C32, S2], BN, ReLU	32xH/2xW/2
(3)	[Conv3x3, C32, S1], BN, ReLU	32xH/2xW/2
(4)	[Conv3x3, C32, S2], BN, ReLU	32xH/4xW/4
(5)	[Conv3x3, C32, S1], BN, ReLU	32xH/4xW/4
(6)	[Conv3x3, C32, S2], BN, ReLU	32xH/8xW/8
(7)	[Conv3x3, C32, S1], BN, ReLU	32xH/8xW/8
	Super-Resolution Module: k=1,2,3,$s = 2^{4-k}$	
(8)	From d^{k-1}: Conv-Sobel, BN	2xH/sxW/s
(9)	[Conv3x3, C32, S1], BN, ReLU	32xH/sxW/s
(10)	From (3/5/7),(9): Concat	64xH/sxW/s
(a)	[Conv3x3, C32, S1, Dil], BN, ReLU	
(b)	ReLU{[Conv3x3, C32, S1, Dil], BN+(a)}	32xH/sxW/s
(11)	Repeat (a-b) with Dil=1,2,4,8,1,1	
(12)	[DConv3x3, C32, S2], BN, ReLU	32x2H/sx2W/s
(13)	[Conv3x3, C1, S1]	1x2H/sx2W/s
(14)	From d^{k-1}: Bilinearly Upsample	1x2H/sx2W/s
(15)	Output d^k: ReLU{(13)+(14)}	1x2H/sx2W/s

performs high frequency residual learning at the resolution level of the input disparity map, and then completes upsampling using the deconvolution layer placed at the last layer of the network. Thus, the computation time of the super-resolution subnetwork is proportional to the input disparity map resolution, rather than the output disparity map resolution. For example, the residual pyramid [23] and the hierarchical refinement [16] structures are all learned in the latter way. This delayed upsampling strategy allows the entire network to operate efficiently in real-time.

More specifically, the operation of extracting high frequency information is defined as follows:

$$\nabla(d) = conv2d(d, w) \tag{4}$$

where the $conv2d$ represents the 2D convolution and the w denotes the Sobel kernel:

$$w_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} w_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{5}$$

Applying the operation defined by (4) to each channel of the left color image, a total of six intensity edge maps can be obtained. The intensity edge maps are sent into three sets of downsample blocks to obtain a pyramid of intensity edge features. Each set of the downsample block contains two 3×3 convolution layers. Similarly, two depth edge maps can be obtained by applying (4) to the input disparity map for each super-resolution scale. 32 features of the depth edge maps are extracted and then stacked with the intensity edge features of the corresponding resolution level. The stacked features are first filtered with a convolution kernel of 32 channels, and then sent to six dilation residual convolution blocks for further depth edges nonlinear mapping. The dilation residual convolution block consists of two 3×3 convolutional layers with a specified dilation factor. We set the dilation factors as $[1, 2, 4, 8, 1, 1]$ to incorporate multiscale contextual information. The last layer is a deconvolution layer which upsamples and aggregates the learned features with a set of deconvolution filters. For every super-resolution scale, we set the deconvolution stride as 2 to avoid introducing obvious checkerboard artifacts. Finally, the output of the super-resolution subnetwork is a one dimension high frequency residuals that is then added to the bilinear upsampled low resolution disparity map.

C. DEPTH DISCONTINUITY AWARE LOSS

In recent years, researchers have trained a variety of deep stereo networks in a fully supervised manner using large scale labeled depth data. Most architecture designs directly minimize depth value differences between the predicted disparity maps and the groundtruth disparity maps, defined by:

$$L_r = \frac{1}{N} \sum_{n=1}^{N} L\left(\hat{d}_n - d_n\right) \tag{6}$$

$\hat{d}_n$ and d_n are the groundtruth disparity value and predicted disparity value for the valid pixel n, respectively. Typically, l_1

measured regression loss is adopted in the work of [12], [18], [19], [24], which is defined by the l_1 norm $L(x) = \|x\|_1$. Smooth l_1 loss is adopted in the work of [13]–[15], [22], which is defined by $L(x) = 0.5x^2\epsilon(1 - |x|) + (|x| - 0.5)\epsilon(|x| - 1)$, where $\epsilon(\cdot)$ is a 0-1 step function. However, it shows that each disparity pixel is optimized independently and equally weighted. Inner connections between disparity pixels have not been fully considered.

In order to exploit the local disparity pixel patterns for explicit supervision, we propose to calculate the first order derivatives on the output disparity maps. In a given disparity map, overlapped regions of different depths form depth discontinuity edges. These edges describe the boundaries of each disparity region and can be detected using a simple gradient detector. We utilize the Sobel kernel defined by (5) and the 2D convolution operation defined by (4) to find the 3x3 local patterns of the disparity map. By learning from the groundtruth depth discontinuity edges, end-to-end stereo networks can get explicit supervision, which is defined by:

$$L_d = \frac{1}{N}\sum_{n=1}^{N} L\left(\nabla\hat{d}_n - \nabla d_n\right) \tag{7}$$

In practice, we calculate the loss of the first derivatives in the x and y directions, respectively. We minimize the following composite loss function for training the proposed deep stereo network:

$$Loss = \sum_{k=0}^{K} L_r^k + \alpha L_d^k \tag{8}$$

where K denotes the total super-resolution scales, and $k = 0$ denotes the raw disparity map, α controls the weight for the L_d loss. The robust loss function proposed by [36] is adopted, which gives $L(x) = \sqrt{\left(\frac{x}{2}\right)^2 + 1} - 1$. By adopting the above composite loss, we are able to guide the network towards discovering depth discontinuity edges in early training epochs. In later training epochs, the depth discontinuity-aware loss decrease gradually and the network turns to focus on minimizing the depth estimation error in depth flat regions.

IV. EXPERIMENTS

In this section, we evaluate our proposed efficient stereo matching network on the Scene Flow dataset [17] and the KITTI datasets [37], [38]. Datasets and implementation details are described first. Then we show the effectiveness of the improvements through ablation studies and give performance comparison with a number of recent state-of-the-art methods on the stereo matching benchmarks.

A. DATASETS AND EVALUATION METRICS

Scene Flow dataset [17] provides 35454 training and 4370 testing synthetic stereo frames with dense disparity groundtruth. We train our models on the final pass of Scene Flow dataset since it involves more post-processing effects such as motion blur, sunlight glare and gamma curve manipulation. KITTI 2012 [37] provides outdoor driving recordings that comprises 194 training and 195 testing images pairs. KITTI 2015 [38] provides 200 training and 200 testing image pairs. Both KITTI datasets give semi-dense groundtruth disparity maps. For the KITTI 2015 dataset, the training set is split into 180 training image pairs and 20 validation image pairs. For the KITTI 2012 dataset, the training set is split into 180 training image pairs and 14 validation image pairs.

For the Scene Flow dataset, we use end-point error (EPE) as the evaluation metric, which is mean average disparity error in pixels. For the KITTI 2012 dataset, the percentage of pixels with error larger than the specified threshold and EPEs for both non-occluded (Noc) and all (All) pixels are reported. For the KITTI 2015 dataset, the percentage of disparity outliers D1 is reported for background (bg), foreground (fg) and all pixels. The outliers are defined as the pixels whose disparity errors are larger than 3 pixels or 5% groundtruth disparity.

B. IMPLEMANTATION DETAILS

All experiments are implemented with PyTorch. The hardware platform is a desktop PC with Intel Core i7-8700K CPU @ 3.7GHz, 32GB RAM and a single NVIDIA GTX 1080Ti GPU with 11 GB memory. All the proposed models are optimized using the RMSprop algorithm. For the Scene Flow dataset, exponentially-decaying learning rate starts from 0.001. The multiplicative factor of learning rate decay is set to 0.9. The batch size is set 4 to maximize the use of GPU memory. We randomly crop patches of size 960×512 from the original stereo images and then feed them to the network. Image intensities are normalized to the range $[-1, 1]$. The maximum disparity value is set to 192. The training process is iterated 25 epochs for every evaluation. For KITTI2012 and KITTI2015, we fine-tune the network pretrained on the Scene Flow dataset for another 500 epochs. The batch size is set to 2. The initial learning rate is 0.001 and is down-scaled by 10 after 300 epochs. Training image pairs are cropped into 256×512, and testing image pairs are padded into 384×1280. The models with the best validation performance on KITTI 2012 and KITTI 2015 are submit to the KITTI server for test set evaluation.

C. ABLATION STUDIES ON SCENE FLOW

In this subsection, we evaluate different model settings and analyze the effectiveness of the proposed architecture in details. The ablation experiments are conducted on the Scene Flow test set. The architecture with the specified number of low resolution hourglass module is denoted with a suffix **Hr**. The architecture with the specified number of super-resolution module is denoted with a suffix **SR**. If the output resolution of the specified super-resolution module does not match the final resolution, a bilinear upsampling is performed.

Firstly, we compare different choices of the α value in Table 3. Only one low resolution hourglass module is utilized in the cost volume aggregation stage, i.e. Hr1. The speed for using different number of super-resolution modules

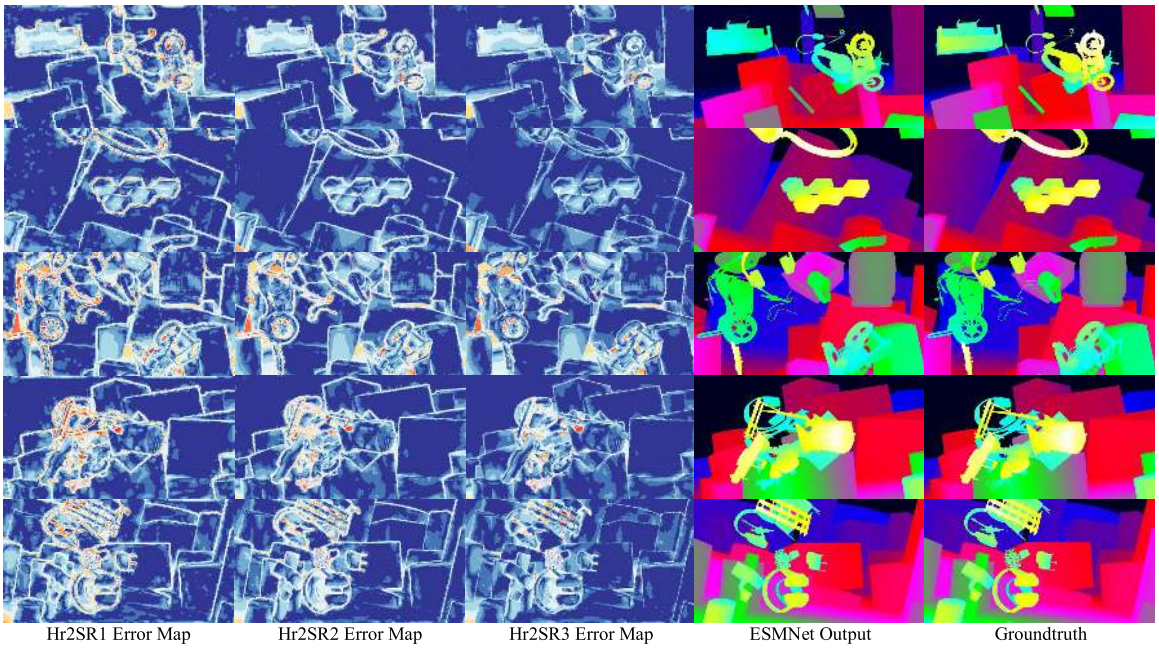

FIGURE 3. Scene Flow test set qualitative results. The error maps of three super-resolution levels are compared. Disparity maps are colored for better visualization.

TABLE 3. Performance comparisons of three models using different α settings. Results are reported on the Scene Flow test set.

α	EPE(px) of ESMNet Variants		
	-Hr1SR1	-Hr1SR2	-Hr1SR3
0	1.431	1.1	0.881
0.15	1.419	1.086	0.875
0.25	1.425	1.088	0.875
0.35	1.419	1.079	0.871
0.45	1.424	1.078	**0.867**
0.55	1.428	1.076	0.872
Time(ms)	35	41	58
FPS	28.6	24.4	17.2

TABLE 4. Retraining performance comparisons of three algorithms using the proposed depth discontinuity aware (DDA) loss. Results are reported on the Scene Flow test set.

Method	EPE(px)			Time (s)
	Reported in Paper	Local Training	Training with DDA Loss	
ESMNet-HR1SR3	N/A	0.881	0.867	0.058
GwcNet [14]	0.765	0.785	0.699	0.36
PSMNet [13]	1.09	1.099	0.99	0.41

is also evaluated. From Table 3 we may find that the best performance is achieved at $\alpha = 0.45$ using the architecture of ESMNet-Hr1SR3. By setting the $\alpha = 0$, it means that the models are trained without the explicit supervision of depth discontinuity edges. Comparing the evaluation performance of $\alpha = 0$ and $\alpha > 0$ in Table 3, it suggests that the proposed depth discontinuity aware loss further reduces the prediction errors of the proposed models. In addition, we have only cascaded up to 3 super-resolution modules, namely SR1, SR2, and SR3, since the total downsampling factor is 2^3. It can be seen from Fig. 3 that by cascading more super-resolution modules, the high frequency details of depth discontinuity edges can be hierarchically recovered, and the depth value noises in depth flat regions can be smoothed, thus, greatly reducing the prediction errors. By timing the evaluation mode of the proposed models, ESMNet-Hr1SR1 achieves a slightly less accurate real-time inference at the 720p resolution level. ESMNet-Hr1SR3 takes 23ms longer than ESMNet-Hr1SR1, but it has achieved state-of-the-art accuracy with near real-time performance (17.2FPS).

As shown in Table 4, we further validate the effectiveness of the proposed depth discontinuity aware (DDA) loss on the other two state-of-the-art stereo matching architectures. We utilize the released codes of GwcNet and PSMNet, and conduct retraining on the local PC. Due to the limited computing resources, both methods use a batch size of 2. Only one GPU is allocated for the retraining of two methods. The α value is set to 0.35 for the retraining of GwcNet. The weights of four output modules set in GwcNet are also applied for calculating the corresponding DDA losses. The retraining process of GwcNet is iterated with the number of epochs specified in the original paper, i.e. 16 epochs. For the PSMNet, the α value is also set to 0.35. We iterate the retraining process of PSMNet with 10 epochs, as specified in the original paper. By retraining using the proposed loss without modifying architecture designs, the prediction accuracy of the selected two deep stereo models, GwcNet and PSMNet, have been further improved by 11% and 9.9%, respectively. Furtherly, we plot the retraining error curves of three models in Fig. 4 to illustrate the effectiveness of the proposed loss. It shows that the error curves of training with DDA loss have a fast decreasing speed in early training epochs. Throughout the retraining process, the proposed DDA loss shows a consistent effect in reducing training errors. The comparison results from Table 4 and Fig. 4 imply that the training

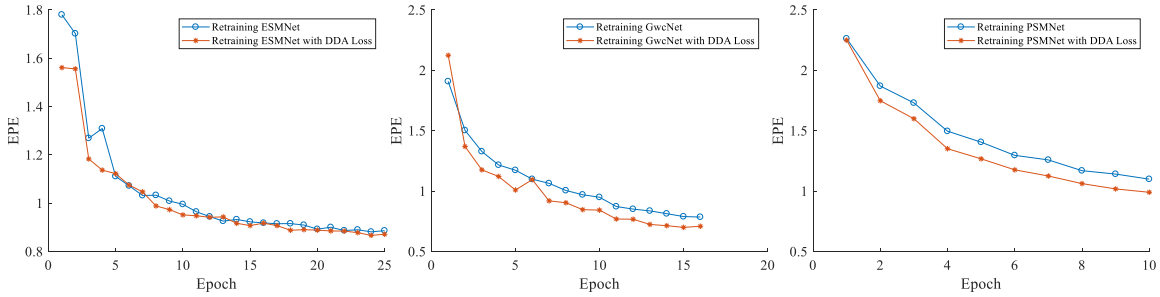

FIGURE 4. Retraining error curves of three methods, i.e. the proposed ESMNet, GwcNet, and PSMNet, respectively. EPE errors are reported on the Scene Flow test set after every training epoch.

TABLE 5. Performance comparisons with state-of-the-art deep stereo matching networks. Results are reported on the Scene Flow test set.

Method	EPE(px)	Num. of 3D Conv. Layers	Time(s)
Other Architectures			
DispNetC [17]	1.68	N/A	0.06
CRL [18]	1.32	N/A	N/A
EdgeStereo [23]	1.11	N/A	0.29
Form 4D Cost Volume and Use 3D Convs.			
GC-Net [12]	2.51	19	0.95
Yu et al. [19]	1.75	19	1.12
StereoNet [16]	1.101	5	N/A
PSMNet [13]	1.09	24	0.41
StereoDRNet [26]	0.86	26	0.28
GA-Net-15 [15]	0.84	15	1.5
GwcNet [14]	0.765	30	0.36
AMNet-32 [22]	0.74	36	1.1
ESMNet-Hr1SR3	0.867	11	0.058
ESMNet-Hr2SR3	0.838	20	0.067
ESMNet-Hr3SR3	0.852	29	0.075

process of stereo matching network can be better guided by mining the intrinsic local patterns between pixels, such as minimizing the differences between the first derivatives of the disparity maps and the first derivatives of the groundtruth disparity maps.

Finally, we compare the performance of three proposed models with 11 state-of-the-art end-to-end stereo matching networks, listed in Table 5. For the architectures following the pipeline of forming 4D cost volumes and applying 3D convolutions, we also list the number of 3D convolution layers used for each model for comparison. As can be seen from the performance of the proposed three models, i.e. from Hr1 to Hr3, increasing the number of 3D convolutional layers can improve the accuracy of stereo matching, but at the cost of a increase in inference time. Since we propose to perform 3D convolution operations at a low resolution level, the added computation time is very limited. It should be noted that the evaluation performance of ESMNet-Hr3SR3 is slightly inferior to the performance of ESMNet-Hr2SR3. This may be because the ESMNet-Hr3SR3 model (3.2M parameters) has more parameters than the ESMNet-Hr2SR3 model (2.4M parameters), requiring more training epochs for optimal performance. Considering the inference speed and the amount of model parameters, only two low resolution hourglass modules are adopted for the proposed efficient stereo matching network. Additionally, our proposed models are much faster than the existing stereo matching algorithms under comparable inference accuracy. For example, the proposed ESMNet-Hr1SR3 are nearly 5 times faster than the recent StereoDRNet [26] method under similar prediction accuracy.

D. PERFORMANCE ON KITTI DATASETS

After training on Scene Flow dataset, we fine-tune the ESMNet-Hr2SR3 model on the KITTI 2015 and KITTI 2012 datasets, respectively. Then we submit the fine-tuned model to the KITTI website for test set evaluation. For these two datasets, we compare the proposed model with 12 recent deep stereo matching models, namely MC-CNN-acrt [6] (CVPR2015), LRCR [39] (CVPR2018), Yu *et al.* [19] (AAAI2018), AMNet [22] (2019), GC-Net [12] (ICCV2017), EdgeStereo [23] (ACCV2018), CRL [18] (ICCV2017), PSMNet [13](CVPR2018), GwcNet [14] (CVPR2019), DispNetC [17] (CVPR2016), StereoNet [16] (ECCV2018) and MADNet [25] (CVPR2019). The comparison results are listed in Table 6 and Table 7. To highlight the efficiency of the proposed architecture, we sort the tables in descending order according to the average running time spent for one testing image pair.

From the tables we may find that the performance of our ESMNet ranks middle level in terms of prediction accuracy. Our proposed architecture lost necessary details by filtering the cost volume at a relative low resolution level, resulting a relative less competitive performance in comparison with recent state-of-the-art methods. The lack of sufficient groundtruth disparity data also limits the performance of the proposed model, which not only considers the loss of disparity values, but also calculates the loss of first derivative of the disparity values. However, our proposed ESMNet speeds up the inference of existing deep stereo networks, especially for the architectures utilizing a series of 3D convolutional layers at the stage of cost aggregation.

For example, on KITTI 2015 dataset, ESMNet is about 13 times faster than GC-Net with similar generalization performance (2.95%/2.72% vs 2.87%/2.61%) and the number of 3D convolution layers (20 vs 19). Comparing with DispNetC, ESMNet improves about 1.39%/1.33% in D1-all under the same running speed. Moreover, comparing to the baseline real-time architecture StereoNet and the recent real-time

TABLE 6. Performance comparisons with state-of-the-art deep stereo matching networks on the KITTI 2015 test set. All deep stereo networks are sorted in descending order of running time.

Methods	All(%)			Noc(%)			Times(s)
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
MC-CNN-art [6]	2.89	8.88	3.89	2.48	7.64	3.33	67
LRCR [39]	2.55	5.42	3.03	2.23	4.19	2.55	49.2
Yu et al. [19]	2.17	5.46	2.79	2.06	5.32	2.32	1.13
AMNet [22]	1.53	3.43	1.84	1.39	3.20	1.69	0.9
GC-Net [12]	2.21	6.16	2.87	2.02	5.58	2.61	0.9
CRL [18]	2.48	3.59	2.67	2.32	3.12	2.45	0.47
PSMNet [13]	1.86	4.62	2.32	1.71	4.31	2.14	0.41
GwcNet [14]	1.74	3.93	2.11	1.61	3.49	1.92	0.32
EdgeStereo [23]	2.27	4.18	2.59	2.12	3.85	2.40	0.27
DispNetC [17]	4.32	4.41	4.34	4.11	3.72	4.05	0.06
StereoNet [16]	4.30	7.45	4.83	n/a	n/a	n/a	n/a
MADNet [25]	3.75	9.20	4.66	3.45	8.41	4.27	0.02
ESMNet	2.57	4.86	2.95	2.41	4.30	2.72	0.067

TABLE 7. Performance comparisons with state-of-the-art deep stereo matching networks on the KITTI 2012 test set. All deep stereo networks are sorted in descending order of running time.

Methods	>2px(%)			>3px(%)			>5px(%)			Mean Error(px)		Times(s)
	Noc	All	Refl.-All	Noc	All	Refl.-All	Noc	All	Refl.-All	Noc	All	
MC-CNN-art [6]	3.90	5.45	27.58	2.43	3.63	20.70	1.64	2.39	14.89	0.7	0.9	67
Yu et al. [19]	2.68	3.42	N/A	N/A	N/A	N/A	1.63	2.23	N/A	0.6	0.7	1.13
AMNet [22]	2.12	2.71	13.56	1.32	1.73	8.16	0.80	1.06	4.21	0.5	0.5	0.9
GC-Net [12]	2.71	3.46	19.07	1.77	2.30	12.80	1.12	1.46	7.99	0.6	0.7	0.9
PSMNet [13]	2.44	3.01	16.06	1.49	1.89	10.18	0.90	1.15	5.64	0.5	0.6	0.41
GwcNet [14]	2.16	2.71	14.57	1.32	1.70	9.28	0.80	1.03	5.22	0.5	0.5	0.32
DispNetC [17]	7.38	8.11	26.54	4.11	4.65	18.15	2.05	2.39	9.88	0.9	1.0	0.06
ESMNet	3.65	4.30	17.86	2.08	2.53	11.11	1.11	1.41	5.89	0.6	0.7	0.067

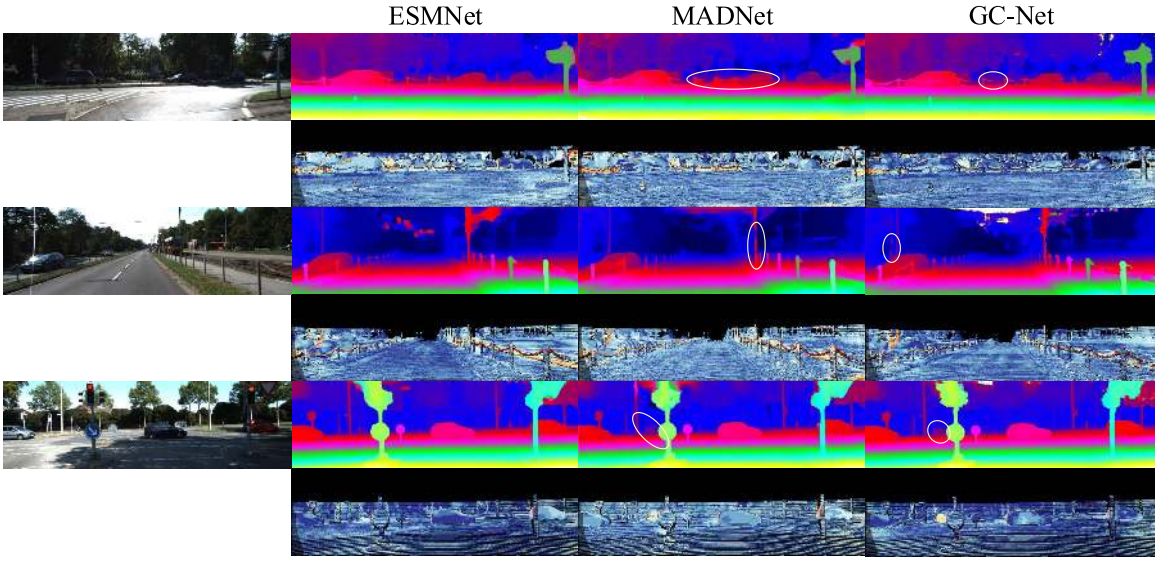

FIGURE 5. KITTI 2015 test set qualitative results. The left column shows the left input image of the stereo image pair. For each input image, the disparity maps obtained by the proposed ESMNet, the recent MADNet [25] and GC-Net [12] are illustrated above corresponding error maps.

architecture MADNet, our ESMNet achieves a relative large performance gain of 38.9% and 36.7%/36.3% in terms of D1-all, while runs in comparable speed. This further proves that the proposed depth discontinuity aware subnetwork learns a much effective hierarchical super-resolution function that runs comparable speed with recent real-time architecture designs.

As for the KITTI 2012 dataset, we cannot find the entries of LRCR, CRL, EdgeStereo, StereoNet and MADNet on the KITTI 2012 leaderboard. Comparing with the listed recent state-of-the-art methods, our methods also gives a middle level performance in terms of prediction accuracy. The proposed ESMNet surpasses DispNetC with a relative improvement of 47%, 45.6% and 41% in terms of 2 pixel, 3 pixel and 5 pixel thresholds, respectively.

Finally, Fig. 5 and Fig. 6 illustrate some visualization examples of the disparity maps estimated by the proposed ESMNet, MADNet, DispNetC, and GC-Net, together with

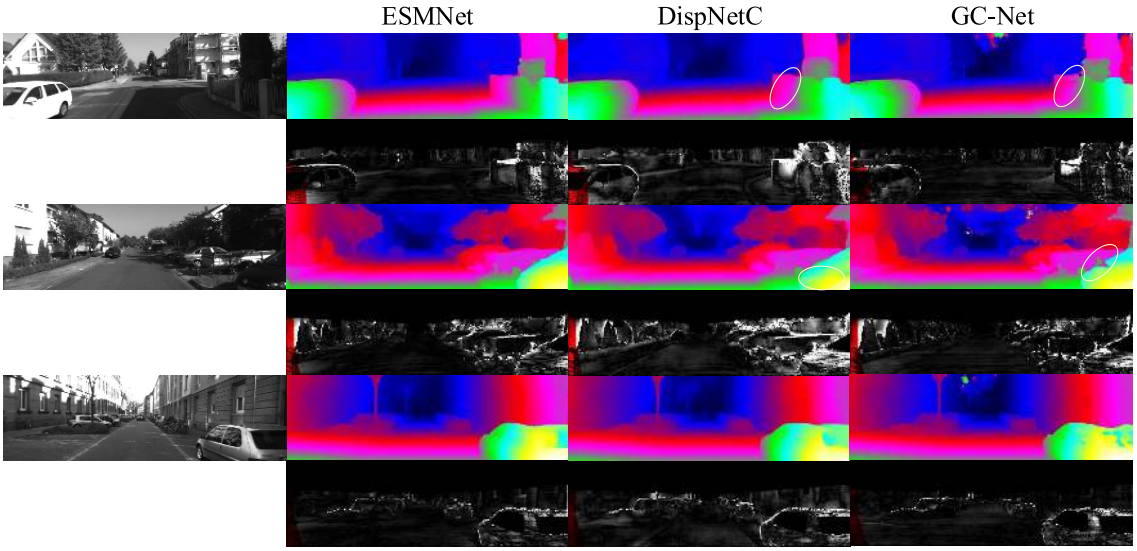

FIGURE 6. KITTI 2012 test set qualitative results. The left column shows the left input image of the stereo image pair. For each input image, the disparity maps obtained by the proposed ESMNet, DispNetC [17] and GC-Net [12] are illustrated above corresponding error maps.

the corresponding error maps. These results were reported by the KITTI server. It shows that ESMNet yields more robust results along depth edges. In general, the proposed ESMNet makes a more practical attempt towards real-time stereo matching with acceptable depth prediction accuracy.

V. CONCLUSION

In this paper, we present ESMNet to address the issue of fast stereo matching by designing a new end-to-end Siamese convolutional neural network architecture. We follow the classic pipeline of forming a 4D matching cost volume, and applying 3D convolutional filtering operations to perform cost aggregation. In order to alleviate the computation burden of 3D convolution operations on high resolution cost volume, we propose to filter and regress on a low resolution cost volume, and then upsample the low resolution disparity map to the desired resolution via the depth discontinuity aware super-resolution subnetwork. We regard the super-resolution process of low resolution disparity as a high frequency cascade refinement process, which explicitly learns a residual mapping function from intensity edges to depth edges. We also propose to supervise the first derivative loss of the predicted disparity map that makes the network adaptively aware the depth discontinuity edges. Experiments on stereo matching datasets verify that ESMNet is capable of predicting high resolution disparity maps at a near real-time frame rate and has a comparable accuracy comparing to the state-of-the-art methods.

REFERENCES

[1] L. Matthies, R. Brockers, Y. Kuwata, and S. Weiss, "Stereo vision-based obstacle avoidance for micro air vehicles using disparity space," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May/Jun. 2014, pp. 3242–3249.

[2] S. Orts-Escolano *et al.*, "Holoportation: Virtual 3d teleportation in real-time," in *Proc. 29th Annu. Symp. Interface Softw. Technol.*, 2016, pp. 741–754.

[3] Y.-J. Lee and M.-W. Park, "3D tracking of multiple onsite workers based on stereo vision," *Autom. Construct.*, vol. 98, pp. 146–159, 2019.

[4] C. Guindel, D. Martín, and J. M. Armingol, "Traffic scene awareness for intelligent vehicles using ConvNets and stereo vision," *Robot. Auton. Syst.*, vol. 112, pp. 109–122, Feb. 2019.

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

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

[7] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.

[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.*, Jun. 2017, pp. 4641–4650.

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

[10] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, "A deep visual correspondence embedding model for stereo matching costs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 972–980.

[11] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 807–814.

[12] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 66–75.

[13] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[14] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3273–3282.

[15] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 185–194.

[16] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "StereoNet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 573–590.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[18] J. Pang, W. Sun, J. S. J. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. ICCV Workshop Geometry Meets Deep Learn.*, Oct. 2017, pp. 887–895.

[19] L. Yu, Y. Wang, Y. Wu, and Y. Jia, "Deep stereo matching with explicit cost aggregation sub-architecture," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7517–7524.

[20] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (PDS): Toward applications-friendly deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018, pp. 5871–5881.

[21] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 979–988.

[22] X. Du, M. El-Khamy, and J. Lee, "AMNet: Deep atrous multiscale stereo disparity estimation networks," 2019, *arXiv:1904.09099*. [Online]. Available: https://arxiv.org/abs/1904.09099

[23] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 20–35.

[24] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "SegStereo: Exploiting semantic information for disparity estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 636–651.

[25] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. Di Stefano, "Real-time self-adaptive deep stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 195–204.

[26] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "StereoDRNet: Dilated residual stereonet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11786–11795.

[27] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[28] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9906, Sep. 2016, pp. 391–407.

[29] G. Riegler, M. Rüther, and H. Bischof, "ATGV-NET: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 268–284.

[30] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 353–369.

[31] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang, and Y. Zhao, "Simultaneous color-depth super-resolution with conditional generative adversarial networks," *Pattern Recognit.*, vol. 88, pp. 356–369, Apr. 2019.

[32] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, and X. Fan, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 5, pp. 26666–26672, 2017.

[33] X. Song, Y. Dai, and X. Qin, "Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2017, pp. 360–376.

[34] W. Zhou, X. Li, and D. Reynolds, "Guided deep network for depth map super-resolution: How much can color help?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1457–1461.

[35] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng, "Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 994–1006, Feb. 2019.

[36] J. T. Barron, "A general and adaptive robust loss function," 2017, *arXiv:1701.03077*. [Online]. Available: https://arxiv.org/abs/1701.03077

[37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3354–3361.

[38] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," in *Proc. ISPRS Workshop Image Sequence Anal. (ISA)*, 2015, pp. 427–434.

[39] Z. Jie, P. Wang, Y. Ling, B. Zhao, Y. Wei, J. Feng, and W. Liu, "Left-right comparative recurrent model for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3838–3846.

CHENGGANG GUO received the M.S. degree in communication and information systems from the South China University of Technology, Guangzhou, China, in 2015. He is currently pursuing the Ph.D. degree in computer application technology with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include visual tracking and depth sensing in augmented reality applications.

DONGYI CHEN received the M.S. and Ph.D. degrees in computer and automation and electronic information engineering from Chongqing University, in 1985 and 1997, respectively. He was a Postdoctoral Research with the Department of Electrical and Computer Engineering, University of Toronto, from 1997 to 1999. He was a Visiting Professor with the School of Computing, Georgia Institute of Technology, from 2002 to 2005. He is currently a Professor with the School of Automation Engineering, University of Electronic Science and Technology of China. His research interests include augmented reality, wearable computing, and wireless sensor networks.

ZHIQI HUANG received the Ph.D. degree in measuring and testing technologies and instruments from the University of Electronic Science and Technology of China, in 2009. He is currently an Associate Professor with the University of Electronic Science and Technology of China. His current research interests include wearable computing, human–computer interaction, and reliability engineering.