IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 4, 2020, accepted July 31, 2020, date of publication August 12, 2020, date of current version August 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016008

Leveraging Contextual Information for Monocular Depth Estimation

DOYEON KIM[1], **(Member, IEEE), SIHAENG LEE**[2], **(Member, IEEE), JANGHYEON LEE**[1], **(Member, IEEE), AND JUNMO KIM**[1,2], **(Member, IEEE)**

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea
[2]Division of Future Vehicle, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

Corresponding author: Junmo Kim (junmo.kim@kaist.ac.kr)

This work was supported by SK Hynix Inc., through the Next Generation Artificial Intelligence Semiconductor System Research Center.

ABSTRACT Humans strongly rely on visual cues to understand scenes such as segmenting, detecting objects, or measuring the distance from nearby objects. Recent studies suggest that deep neural networks can take advantage of contextual representation for the estimation of a depth map for a given image. Therefore, focusing on the scene context can be beneficial for successful depth estimation. In this study, a novel network architecture is proposed to improve the performance by leveraging the contextual information for monocular depth estimation. We introduce a depth prediction network with the proposed attentive skip connection and a global context module, to obtain meaningful semantic features and enhance the performance of the model. Furthermore, our model is validated through several experiments on the KITTI and NYU Depth V2 datasets. The experimental results demonstrate the effectiveness of the proposed network, which achieves a state-of-the-art monocular depth estimation performance while maintaining a high running speed.

INDEX TERMS Monocular depth estimation, contextual information.

I. INTRODUCTION

Depth estimation is a key problem in computer vision that can be applied to a variety of fields such as autonomous driving, 3D modeling, or robotics. In particular, monocular depth estimation aims to generate a corresponding depth map for a given image, which is an ill-posed task. This is because a number of distinct 3D scenes can be mapped to a single 2D image. However, humans can estimate the distance to objects even with one eye because they can exploit semantic features [1] and monocular cues. Recent papers support that convolutional neural networks (CNNs) also take advantage of a similar property. Hu *et al*. [2] trained an auxiliary mask network that can predict the minimum set of relevant pixels in the image that can contribute to the inference of the depth map. Through visualization of the predicted mask, they have found that CNNs can use visual cues, such as edges or boundaries in input images, and inside the region of individual objects. This study indicates that semantic features can play a crucial role in depth estimation for humans and deep neural networks. Hence, focusing on the contextual information in input images can be beneficial for effective monocular depth estimation.

Since the emergence of the deep neural networks, there has been a rapid rise in the state-of-the-art performance in monocular depth estimation. By adopting a good backbone network trained on a substantially large-scale dataset, it became easier to extract more powerful features. Thus, many researchers have studied methods for applying the knowledge acquired from this powerful encoder for depth estimation [3]–[6]. Moreover, several papers have attempted to leverage contextual features in this area. Reference [7], [8] employed an encoder–decoder structure with a skip connection; however, their methods focus more on refining the coarse local features than contextual information itself. Some studies have used additional knowledge such as pretrained weights or a segmentation dataset [9] to achieve semantic supervision; however, these methods limit the datasets that can be applied and it complicates the training methodology. Therefore, it is worth formulating a training strategy that allows the network to concentrate on significant regions and uses semantic representations without employing any external information.

This paper proposes a new network architecture to leverage the contextual information for effective monocular depth estimation. The first contribution of this paper is the proposed attentive skip connection which enables the use of encoded features in the decoding phase. As previously discussed, objects with different positions and sizes can play a crucial

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

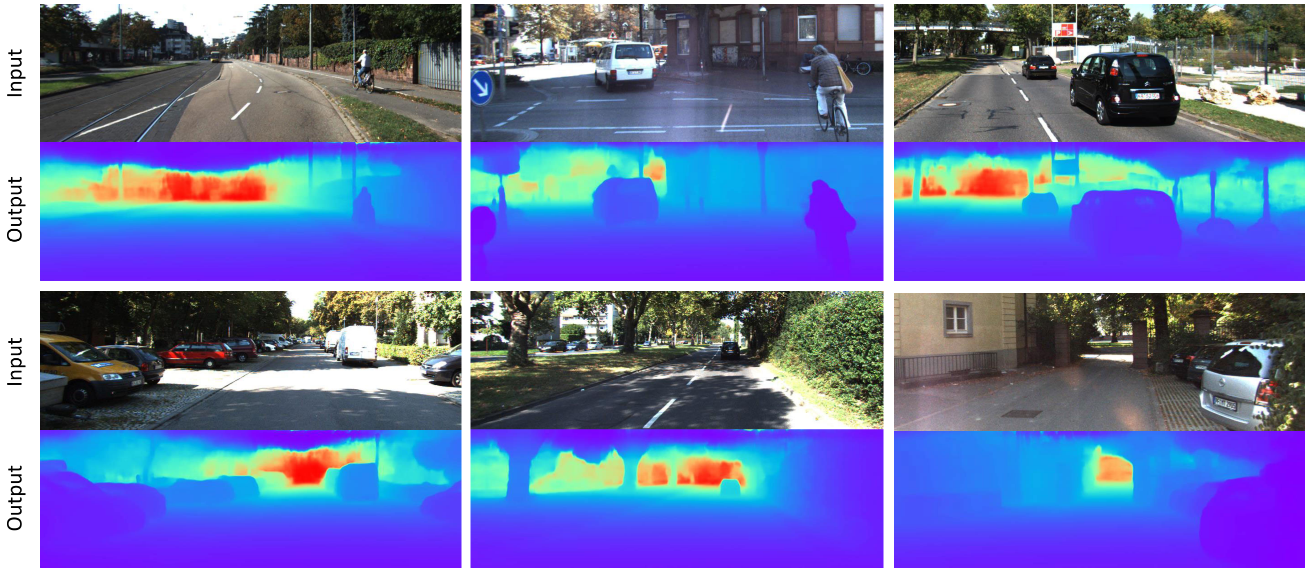

FIGURE 1. Generated depth maps on the KITTI dataset. The first and third rows are the input RGB images. The second and fourth rows are visualized with the depth maps from the input images.

role in depth estimation. Therefore, a multi-scale skip connection with self-attentive modules is added to highlight the feature maps from the diverse objects in a different scale. The second contribution of this paper is a novel global context module, which leverages global features to understand the scene context comprehensively in a global scale. The global context module receives the bottleneck feature as an input and captures rich contextual information. These additional units are adoptive for all networks and it consumes a small amount of computation, which yields a high inference speed. By focusing on significant regions and representation with effective light weight augmented modules, the model shows a high performance with a reasonable inference time. To summarize, the main contributions of this study are as follows:

- Contextual information plays an important role in many scene understanding tasks, including monocular depth estimation. To generate an accurate depth map for a given image, we introduce a novel network architecture that leverages contextual information using an encoder–decoder structure.
- The novel attentive skip connection delivers the features that are obtained from the encoder to the decoder; hence, the model can take advantage of the encoded features in the decoding phase. In contrast with previous studies involving skip connections [10], an attentive skip connection infers an attention map to learn the regions on which the network should focus.
- We propose a global context module to enhance the obtained bottleneck feature and to exploit the global context for a comprehensive scene understanding.
- The experimental results demonstrate that the proposed model accomplishes a state-of-the-art performance on the KITTI and NYU depth V2 datasets. Owing to the easy integration of the lightweight modules, the network shows a high running speed while improving the performance in comparison to previous methods.

II. RELATED WORK

A. MONOCULAR DEPTH ESTIMATION

There has been significant development in monocular depth estimation. Wang *et al.* [11] solved semantic segmentation and depth estimation tasks jointly by developing a unified framework. They employed joint global and regional CNNs to predict potential and inferred final results through the hierarchical conditional random field. Laina *et al.* [12] proposed fully convolutional networks with the fast up-projection method using residual learning to model the mapping between RGB images and depth maps. Furthermore, they introduced the reverse Huber loss, which tackles the heavy-tailed distribution of the depth dataset. Godard *et al.* [13] suggested unsupervised training objective to replace the use of labeled depth maps. The network generates the left and right disparity maps and calculates the reconstruction, smoothness, and left-right consistency terms. Kuznietsov *et al.* [14] introduced a semi-supervised approach to overcome the deficiency and limitation of sparse ground truth lidar maps. They trained the network with sparse depth maps in a supervised manner and provided image alignment loss to generate photoconsistent dense maps based on stereo images. Li *et al.* [15] showed the two-streamed network that produces depth and depth gradients with the given RGB image and combines each result to obtain a final dense depth map. Fu *et al.* [3] modeled the monocular depth estimation as a classification task and tackled this problem with a spacing-increasing discretization strategy. Gan *et al.* [5] employed an affinity layer to integrate relative and absolute features within a network. In addition, they used vertical max pooling to focus on vertical characteristics of depth maps and improved accuracy.

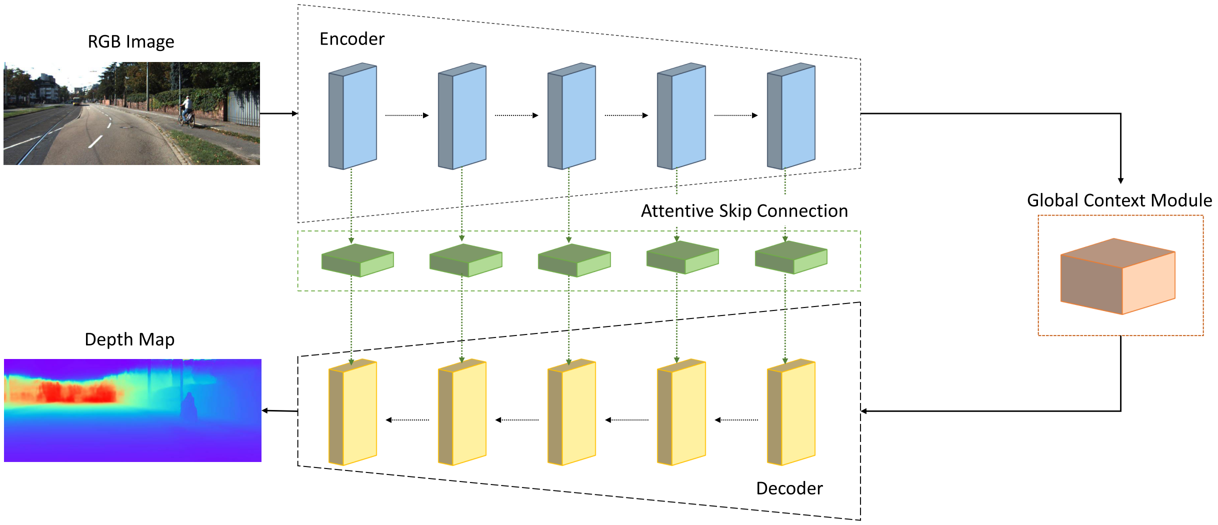

FIGURE 2. The entire architecture of the proposed monocular depth estimation network. It consists of an encoder–decoder network with the proposed attentive skip connection (ASC) and the global context module (GCM) for effective depth prediction. The ASC (green blocks and lines) is located between each encoder and the decoder block. The GCM (orange block) is placed between the encoder and the decoder network.

Guo *et al.* [6] incorporated a synthetic depth dataset to acquire a considerable amount of ground truth images. Subsequently, they trained a network with synthetic data and fine-tuned with a real dataset. Finally, they mitigated the domain gap between the ground truth and synthetic dataset by distilling stereo networks. Qi *et al.* [16] utilized the relation between the depth and surface normal by employing two networks: depth-to-normal and normal-to-depth networks. Hu *et al.* [17] proposed a network that extracts a multi-scale feature to preserve spatial resolution. In addition, they defined a new loss that considers the depth, gradients, and surface normal of depth maps. Yin *et al.* [4] emphasized the importance of geometric constraints in the 3D space to improve the performance of monocular depth estimation. They generated a 3D point cloud from the estimated and ground truth depth maps, and followed by computing the virtual normal loss by randomly sampling points of pair maps. Zhang *et al.* [18] suggested a new framework that predicts depth, surface normal, and semantic segmentation jointly. This framework utilizes cross-task patterns by calculating the affinity matrix while performing each task.

B. CONTEXTUAL INFORMATION

Contextual information is an essential cue in many computer vision tasks, especially in scene understanding tasks such as 3D object detection, semantic segmentation, or depth estimation. Reference [10] constructed an encoder–decoder architecture with skip connections to combine contracted high-resolution features with an expanded output for segmentation. To achieve depth estimation, Eigen *et al.* [7] and Garg *et al.* [8] employed the encoder–decoder structure with skip connections that use encoded features in the decoding phase. Liu *et al.* [19] suggested a network that performs semantic segmentation first and uses the predicted labels for depth estimation. Jiao *et al.* [20] proposed a synergy network to incorporate semantics in depth prediction by using an information propagation strategy as well as knowledge sharing. Amirkolaee and Arefi [21] constructed a depth prediction network with the encoder–decoder and skip connection structure to integrate the global and local contexts. Unsupervised methods use additional information to overcome the absence of labeled data; such methods include those that leverage semantic information. Ochs *et al.* [9] performed semantic segmentation and depth estimation using two independent CNNs, one for each task. Through this approach, the network learns more stable features and can leverage semantic labels. Chen *et al.* [22] combined the depth and segmentation modalities by minimizing self-supervised objective losses, the left–right semantic consistency, and the semantics-guided disparity smoothness.

In addition, there have been several papers that have employed an attention architecture to focus on the contexts that are significant for depth estimation. Xu *et al.* [23] proposed an attention module which parameterized by binary variables to control the flow between the encoder and the decoder. Then, the proposed attention module was integrated with a conditional random field. Chen *et al.* [24] proposed an attention-based context aggregation network to solve the depth estimation problem. They placed a pixel-level self-attention module at the bottleneck of network and trained it with attention loss. Takagi *et al.* [25] proposed a two-branch depth estimation network with mutual learning and employed channel attention with squeeze–and– excitement [26] attention module.

III. METHOD

This section first introduces the entire architecture of the network and an attentive skip connection with the global context module in the sequence.

A. NETWORK ARCHITECTURE

As depicted in Fig. 2, the proposed model adopts the encoder–decoder architecture with the suggested attentive skip connections and the global context module. The encoder is initialized with the weights of a pretrained ImageNet [27] classification model to extract the dense features. We develop a remarkably simple decoder network to restore the obtained features to the image scale and to generate a depth map. The decoder is designed to have the same number of blocks as the encoder. Each block consists of a 3 × 3 deconvolution, batch normalization, and ReLU layer. To strengthen the power of the decoder, residual blocks are placed between the 2^{nd} and 3^{rd} block. Similar to the decoder blocks, the residual blocks have two 3 × 3 convolution layers with batch normalization and a ReLU layer.

B. ATTENTIVE SKIP CONNECTION

We consider a depth estimation network as a mapping function for the image to depth map translation, which shares an underlying structure. The objects and structure in a given RGB image are roughly aligned with those in the output depth map. As previously described, the location of important edges plays a major role in depth inference. Therefore, it would be desirable to flow acquired information through the network. In this study, to shuttle the low-level features, we append the skip connections between the encoder and the decoder. Unlike previous studies [4], [7], [8], this study does not simply sum the feature values or apply a concise convolution. An attention mechanism is applied to the skip connection. As discussed earlier, there are some studies that have used the attention mechanism for monocular depth estimation [23]–[25]. However, our approach differs from previous works in two aspects. First, we design a task–specific attention module with two branches and attach it to the skip connection to deliver refined multi-scale features to each of the blocks of the decoder. Second, our module is light-weight and requires only a small amount of additional computation. The proposed attention module is detailed in Fig. 3.

An attentive skip connection is provided for each encoder block to propagate the meaningful features to the decoder block. For every convolutional block in the encoder, the output feature maps F_M pass through two branches. Similar to the implementation in [28], the attentive skip connections generate attention maps along the spatial and channel dimensions. In the first branch, a spatial attention map is obtained through its branch in parallel with the channel attention branch. We consider that the computational graph for the spatial attention map should be task-specific. Previously proposed attention modules are usually employed for the classification task [26], [28], [29]. It is needed to derive the highest possibility from the whole image in the classification task; however, the regression network infers continuous values for all of the pixels in the image. Therefore, in this study, an attentive skip connection is designed to specifically understand the scene in multiple scales. As it is necessary for our network to focus on multiple locations, important edges, and objects during depth estimation. We adopt atrous spatial pyramid pooling (ASPP) to broaden the fields-of-view and to capture the objects at multiple scales. The intermediate feature map F_M for each block of the encoder is forwarded to the ASPP module with the dilation rate $d = \{d_1, d_2, \cdots, d_n\}$. The value of d is obtained via experiments. We choose {3, 6, 9} for this investigation and this process will be discussed in the Experiment section. Furthermore, the feature that passed the ASPP module goes through a 1 × 1 convolution for effective integration. The integrated feature are passed through the ASPP module and the 1 × 1 convolution once more to enhance the effectiveness of the module. Ultimately, a batch normalization layer is employed at the end of the spatial block to ensure stable training. Thus, the spatial attention map F'_M is acquired.

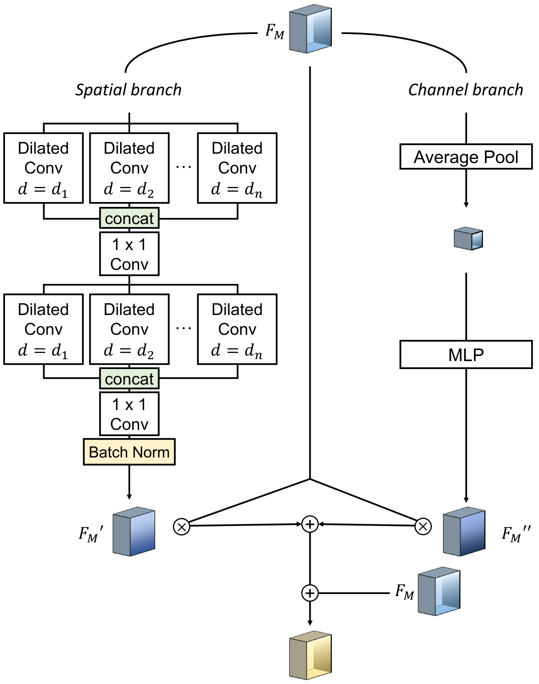

FIGURE 3. Detailed structure of an attentive skip connection. F_M is an intermediate output feature from the encoder. F'_M, F''_M are attention maps from each spatial and channel branch. The dilation rate of the atrous spatial pyramid pooling (ASPP) module in spatial branch $d = \{d_1, d_2, \cdots d_n\}$ is obtained through experiment.

In the second branch, average-pooling is applied for the intermediate feature map F_M in the channel dimension to encode the contextual information in each channel. Then, the pooled feature is forwarded to a multi-layer perceptron (MLP) with one hidden layer. To make the model compact and effective, a hidden layer is constructed to have a reduced number of units compared to that of the input and output layers of the MLP. The value of 16 is used as the reduction ratio for the dimensions of the hidden and input layers. Thus, a refined spatial attention map, F''_M, is obtained. After the spatial attention map F'_M and the channel attention map F''_M are acquired, each map is multiplied with the original feature map F_M element-wisely and they are merged by summation. Finally, the calculated refined feature is added with the original feature map F_M.

C. GLOBAL CONTEXT MODULE

To further exploit a global context representation, we do not directly deliver the bottleneck feature of the encoder to

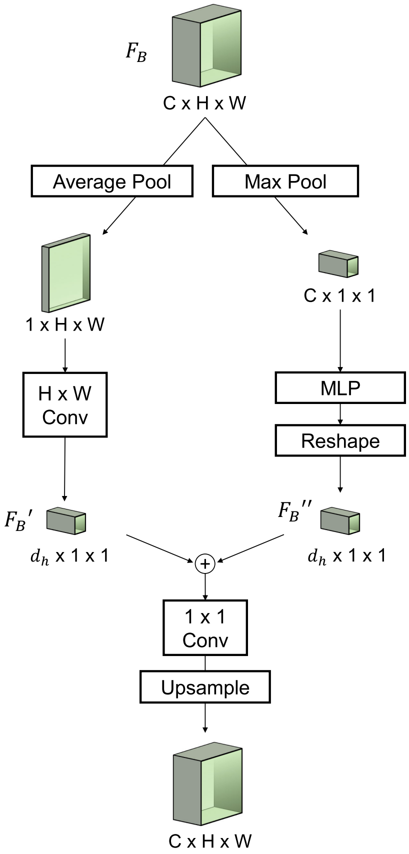

FIGURE 4. **Illustration of the global context module. F_B is the bottleneck feature from the encoder, with a size of $C \times H \times W$. F_B' and F_B'' denote the processed feature from each branch and d_h is the hidden dimension of the branch. We determine the optimal value of d_h through additional experiments.**

the decoder. The global context module is placed at the end of the encoder to obtain the global context information and pass meaningful features to the decoder. The structure of the global context module is illustrated in Fig. 4. The bottleneck feature $F_B \in \mathbb{R}^{C \times H \times W}$ is fed into two paths. The goal of the global context module is to capture important features in a global scale with a simple additional computation. Hence, the pooling method is applied to reduce the dimensions of the feature and obtain significant representations with small parameter overhead. In the first branch, average pooling is applied in the channel dimension to utilize the inter-dependencies between the channel-wise feature maps and to help the model concentrate on the useful regions since average-pooling has been commonly used for capturing spatial information [26]. Then, the feature is convolved with the kernel having $H \times W \times d_h$ weights where d_h denotes the dimension of the intermediate refined feature map $F_B{}'$. The appropriate dimension for the best performance is determined to be 512 via ablative experiments. Regarding the second branch, a max-pooling operation for F_B is used to capture the most informative spatial information. Similar to the case of the attentive skip connection, the max-pooled feature is forwarded to the multi-layer perceptron comprising one hidden layer with a reduced dimension, and the reduction ratio 16. Then, the refined feature is reshaped into $d_h \times 1 \times 1$ to aggregate it with $F_B{}'$.

After the refined feature maps $F_B{}'$ and $F_B{}''$ are obtained from both branches, the output vectors are combined using element-wise summation. Additionally, a 1×1 convolution layer is employed to fuse the added features, and it is upsampled by bilinear interpolation such that it has the same size as that of the original feature map F_B. Finally, this obtained feature is multiplied and added with the original feature map F_B and used as an input to the decoder.

D. TRAINING

In the training phase, a scale-invariant log loss function [7] is used as the objective function. For a generated depth map y and the ground truth y^*, there are n pixels indexed by i. The final loss function is as follows:

$$L_{obj}(y, y^*) = \frac{1}{n}\sum_i {d_i}^2 - \frac{1}{2n^2}\left(\sum_i {d_i}^2\right) \tag{1}$$

where $d_i = \log y_i - \log {y_i}^*$.

IV. EXPERIMENT

The effectiveness of the proposed model is demonstrated by performing various experiments on the KITTI [36] and NYU Depth V2 [37] datasets. For the evaluation, this study uses the following metrics from previous works [3], [7]:

- Threshold (δ):

$$\% \text{ of } y_i \text{ s.t. } max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thr$$

- Absolute relative difference (AbsRel):

$$\frac{1}{|T|}\sum_{y \in T} |y - y^*|/y^*$$

- Squared relative difference (SqRel):

$$\frac{1}{|T|}\sum_{y \in T} \|y - y^*\|^2/y^*$$

- Root mean squared error (RMSE):

$$\sqrt{\frac{1}{|T|}\sum_{y \in T} \|y - y^*\|^2}$$

- RMSE (log):

$$\sqrt{\frac{1}{|T|}\sum_{y \in T} \|\log y - \log y^*\|^2}$$

- log10:

$$\frac{1}{|T|}\sum_{y \in T} |\log_{10} y - \log_{10} y^*|$$

where T is the available pixels in the ground truth, y is the predicted value, and y^* is the ground truth. Following the illustration of our results on the dataset, an ablation study has been provided.

TABLE 1. Performance on the KITTI Dataset.

Method	Base Network	cap	*higher is better*			*lower is better*			
			δ_1	δ_2	δ_3	AbsRel	SqRel	RMSE	RMSE log
Saxena *et al.* [30]	-	0 - 80m	0.601	0.820	0.926	0.280	3.012	8.734	0.361
Liu *et al.* [19]	[31]	0 - 80m	0.647	0.882	0.961	0.217	1.841	6.986	0.289
Eigen *et al.* [7]	-	0 - 80m	0.692	0.899	0.967	0.190	1.515	7.156	0.270
Nath *et al.* [32]	ResNet-50	0 - 80m	0.771	0.922	0.971	0.167	1.257	5.578	0.237
Xu *et al.* [23]	ResNet-50	0 - 80m	0.818	0.954	0.985	0.122	0.897	4.677	-
Godard *et al.* [13] (CS+K)	ResNet-50	0 - 80m	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Kuznietsov *et al.* [14]	ResNet-50	0 - 80m	0.862	0.960	0.986	0.113	0.741	4.621	0.189
Gan *et al.* [5]	ResNet-50	0 - 80m	0.890	0.964	0.985	0.098	0.666	3.933	0.173
Guo *et al.* [6]	VGG-16	0 - 80m	0.892	0.967	0.986	0.096	0.641	4.095	0.168
Chen *et al.* [24]	ResNet-101	0 - 80m	0.919	0.982	0.995	0.083	0.437	3.599	0.127
Fu *et al.* [3]	ResNet-101	0 - 80m	0.932	0.984	0.994	0.072	0.307	2.727	0.120
Yin *et al.* [4]	ResNeXt-101	0 - 80m	0.938	0.990	0.998	0.072	-	3.258	0.117
Ours	ResNet-50	0 - 80m	0.950	0.993	0.998	0.068	0.262	2.798	0.130
Ours	ResNet-101	0 - 80m	0.956	0.993	0.999	0.063	0.242	2.690	0.097
Ours	ResNeXt-101	0 - 80m	**0.958**	**0.993**	**0.999**	**0.060**	**0.231**	**2.650**	**0.094**
Garg *et al.* [8]	AlexNet	0 - 50m	0.740	0.904	0.962	0.169	1.080	5.104	0.273
Godard *et al.* [13] (CS+K)	ResNet-50	0 - 50m	0.873	0.954	0.979	0.108	0.657	3.729	0.194
Kuznietsov *et al.* [14]	ResNet-50	0 - 50m	0.875	0.964	0.988	0.108	0.595	3.518	0.179
Gan *et al.* [5]	ResNet-50	0 - 50m	0.898	0.967	0.986	0.094	0.552	3.133	0.165
Guo *et al.* [6]	VGG-16	0 - 50m	0.901	0.971	0.988	0.092	0.515	3.163	0.159
Fu *et al.* [3]	ResNet-101	0 - 50m	0.936	0.985	0.995	0.071	0.268	2.271	0.116
Ours	ResNet-50	0 - 50m	0.944	0.992	0.998	0.069	0.293	3.229	0.107
Ours	ResNet-101	0 - 50m	0.947	0.992	0.999	0.066	0.291	3.189	0.103
Ours	ResNeXt-101	0 - 50m	**0.964**	**0.995**	**0.999**	**0.058**	**0.163**	**1.893**	**0.088**

A. IMPLEMENTATION DETAILS

This model is implemented on the open deep learning framework PyTorch [38]. The encoder is initialized with the weights of the pretrained networks ResNet-50, ResNet-101 [39], ResNeXt-101 [40]. We use randomly cropped images with a size of 352 × 704 from the KITTI dataset and images with a size of 448 × 576 from the NYU Depth V2 dataset. The learning strategy employs the ADAM optimizer, and the learning rate is started from 0.0001 with a weight decay of 0.9. The network is trained for 40 epochs and the batch size is set to four. The images are augmented by applying random brightness, contrast, color adjustment, and rotation; the range for each of the aforementioned modifications is (0.5, 1.5), (0.8, 1.2), (0.8, 1.2), and (-5, 5) degrees, respectively. In addition, random horizontal flipping is applied.

B. DATASET

1) KITTI

The KITTI dataset [36] consists of 61 scenes of outdoor images captured by driving a car with cameras and velodyne sensors. The proposed model is trained based on the split proposed by Eigen *et al.* [7]. They used 56 scenes from the ‘‘city’’, ‘‘residential’’, and ‘‘road’’ categories. The images are split into training and testing sets, which contain 23,488 and 697 images, respectively.

2) NYU DEPTH V2

The NYU Depth V2 dataset [37] contains 464 indoor scenes, which includes 249 scenes for training and 215 for testing. The proposed model is trained on 24,231 images and tested on 654 images.

TABLE 2. Performance on the NYU Depth V2 Dataset.

Method	Base Network	*higher is better*			*lower is better*		
		δ_1	δ_2	δ_3	AbsRel	log_{10}	RMSE
Wang *et al.* [11]	-	0.605	0.890	0.970	0.220	-	0.871
Liu *et al.* [19]	[31]	0.650	0.906	0.976	0.213	0.087	0.759
Eigen *et al.* [33]	VGG-16	0.769	0.950	0.988	0.158	-	0.641
Laina *et al.* [12]	ResNet-50	0.811	0.953	0.988	0.127	0.055	0.573
Li *et al.* [15]	VGG-16	0.789	0.955	0.988	0.152	0.064	0.611
Xu *et al.* [23]	ResNet-50	0.806	0.952	0.986	0.125	0.057	0.593
Chen *et al.* [24]	ResNet-101	0.826	0.964	0.990	0.138	-	0.496
Fu *et al.* [3]	ResNet-101	0.828	0.965	0.992	0.115	0.051	0.509
Amirkolaee *et al.* [21]	ResNet-110	0.830	0.968	0.990	0.115	0.049	0.523
Qi *et al.* [16]	ResNet-50	0.834	0.960	0.990	0.128	0.057	0.569
Hao *et al.* [34]	ResNet-101	0.841	0.966	0.991	0.127	0.053	0.555
Zhang *et al.* [18]	ResNet-50	0.846	0.968	0.994	0.121	-	0.497
Hu *et al.* [17]	SENet-154	0.866	0.975	0.993	0.115	0.050	0.530
Lin *et al.* [35]	SENet-154	0.866	0.975	0.993	0.115	0.050	0.523
Takagi *et al.* [25]	SENet-154	0.873	0.976	0.994	0.113	0.049	0.521
Yin *et al.* [4]	ResNeXt-101	0.875	0.976	0.994	**0.108**	0.048	0.416
Ours	ResNet-50	0.850	0.975	0.994	0.126	0.053	0.423
Ours	ResNet-101	0.867	0.978	0.996	0.118	0.050	0.400
Ours	ResNeXt-101	**0.878**	**0.981**	**0.995**	0.111	**0.047**	**0.388**

C. PERFORMANCE

The results obtained for the KITTI and NYU Depth V2 datasets are listed in Table 1 and Table 2, where the proposed model is compared with other previous works. As described in the results, our approach outperforms the other state-of-the-art methods for the outdoor and indoor datasets. It proves that the proposed model is suitable for various situations. As presented in Table 1, the results of our method exceed those of the previous works by 2% ∼ 22% in terms of all of the metrics on the KITTI dataset. From Table 2, our model achieves state-of-the-art results for all of the metrics, except for AbsRel on the NYU Depth V2 dataset.

In Fig. 5, the results are compared with those of prior works, for the KITTI dataset. As previously highlighted, the importance of contextual information in depth estimation has been demonstrated. The proposed method shows sharp boundaries on objects such as a person, a road sign, or bicycles. In addition, our approach successfully locates the

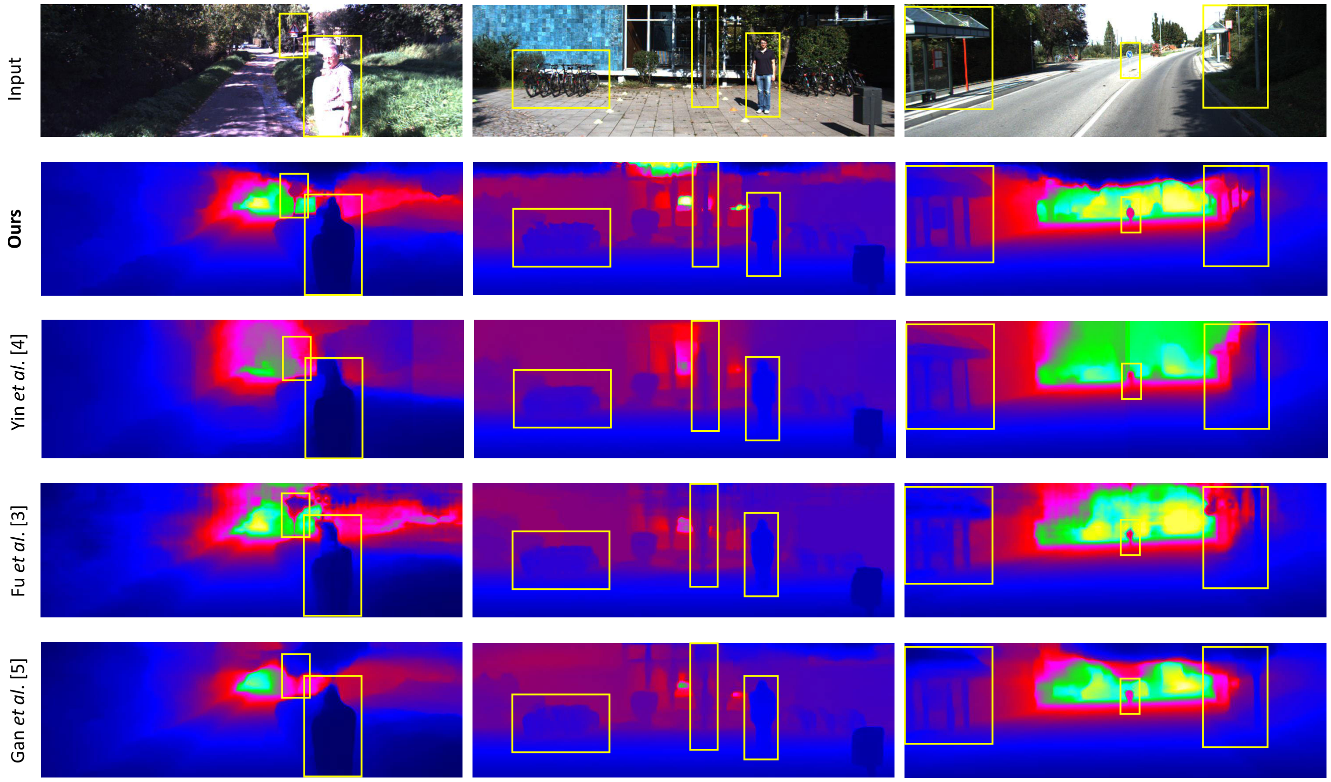

FIGURE 5. Qualitative comparison with the previous methods. The depth maps are generated from the test set of the KITTI dataset. From top to bottom, the images are the input and the depth map of our method and those of the methods propsed by Yin [4], Fu [3], and Gan [5].

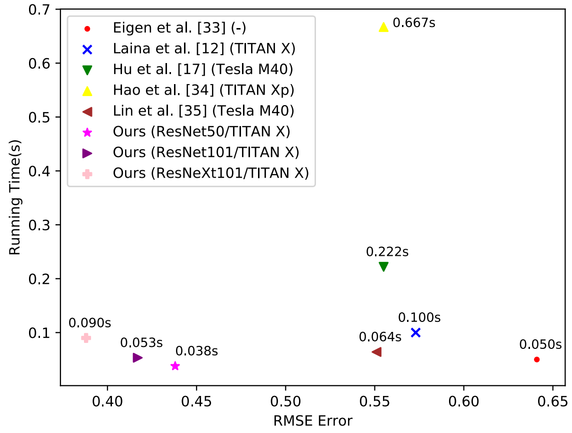

FIGURE 6. Comparison of the inference speed with the previous methods: running time vs. RMSE error. The performance and running time of the other methods are derived from [35].

objects in the image, in contrast with the previous methods, even when there are multiple objects. The road sign in the image in the first column is not presented in the result depth map of [4], [5]; in contrast, our model provides an appropriate inference for the depth of a given object.

To further emphasize the strength of the proposed method, the mean RMSE versus the running time for the proposed model on the NYU Depth V2 dataset is illustrated along with that for some of the prior works. As shown in Fig. 6, the inference speed of our method is higher than that of the other compared methods; in addition, our model achieves a higher accuracy. The results are represented from different base networks including ResNet-50, ResNet-101, and ResNeXt-101. Even though there is a trade-off between the performance and the inference time, the suggested model consistently provides reasonable results.

D. ABLATION STUDY

To demonstrate the effectiveness of the proposed model, we conduct several ablation experiments with different settings on the NYU Depth V2 dataset. First, experiments are performed on the baseline method; then, the network is amended with an attentive skip connection and a global context module to verify the performance of the proposed method. The quantitative and qualitative results are shown in Table 3 and Fig. 7.

As listed in Table 3, the attentive skip connection and the global context module significantly improve the performance of the network. To demonstrate that this strategy can be generallized to a different base network, the model is trained with ResNet-101 and ResNeXt-101. The results shows that the proposed approach consistently exhibits good performance even when applied to a different network. Furthermore, the number of parameters increased only by 2.8M, as listed in the table. A significant improvement in the performance and fast inference are achieved with a small number of parameters for the suggested modules. The qualitative results obtained for the proposed modules on the NYU Depth V2 dataset are illustrated in Fig. 7. It can be observed that the boundaries of the objects became more accurate owing to the addition of the

TABLE 3. Ablation study on the NYU Depth V2 dataset. Baseline: encoder–decoder network with the skip connections; ASC: attentive skip connection; GCM: global context module. The encoder and the decoder are the same for all settings.

Method	Base Network	δ_1	δ_2	δ_3	AbsRel	SqRel	RMSE	SILog	Params
Baseline	ResNet-101	0.837	0.971	0.994	0.131	0.080	0.439	13.142	57.0M
Baseline + ASC		0.859	0.977	0.995	0.118	0.070	0.421	11.958	57.6M
Baseline + ASC + GCM		**0.867**	**0.978**	**0.996**	**0.115**	**0.067**	**0.403**	**11.850**	59.8M
Baseline	ResNeXt-101	0.865	0.977	0.995	0.118	0.071	0.407	12.157	101.2M
Baseline + ASC		0.871	0.979	**0.996**	0.114	0.066	0.390	11.488	101.9M
Baseline + ASC + GCM		**0.878**	**0.981**	0.995	**0.111**	**0.065**	**0.388**	**11.315**	104.0M

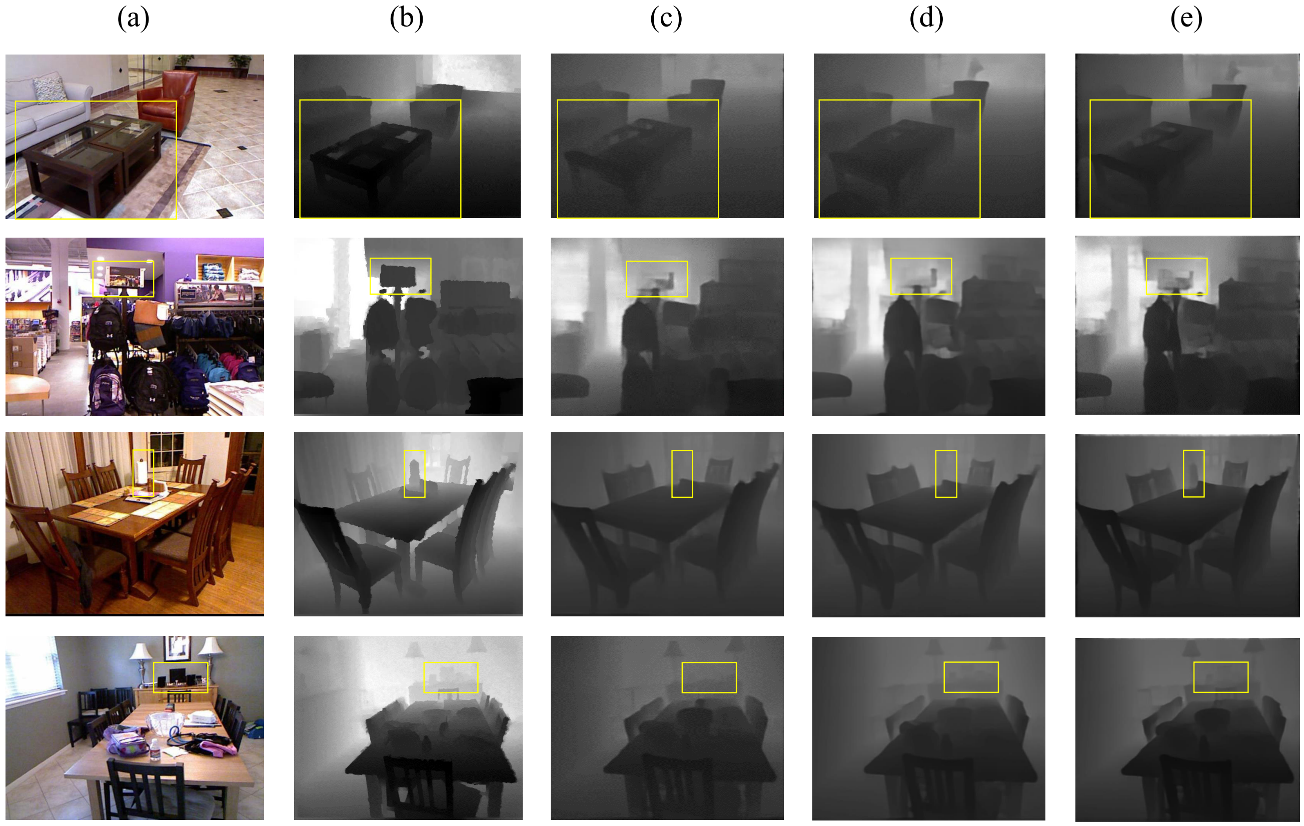

FIGURE 7. Qualitative results of the ablation study. (a) input RGB images; (b) ground truth; (c) baseline; (d) baseline and attentive skip connection (e) baseline, attentive skip connection, and the global context module (ours).

TABLE 4. Comparison with previous attention modules.

Method	AbsRel	SqRel	RMSE	SILog
SE [26]	0.069	0.275	2.780	9.405
BAM [28]	0.063	0.253	2.765	9.024
CBAM [29]	0.069	0.254	2.775	9.239
Ours	**0.061**	**0.246**	**2.754**	**8.930**

proposed modules. Moreover, our model is able to accurately detect the objects on the table (3rd row) that were not detected accurately by the baseline method.

Table 4 shows the results of the comparison of the proposed attentive skip connection with previous attention modules. Squeeze–and–excitement (SE) [26], bottleneck attention module (BAM) [28], and convolutional bottleneck attention module (CBAM) [29] are selected and tested on the KITTI dataset based on ResNet-101 architecture. The proposed attentive skip connection yields the best performance for all metrics, as indicated by Table 4. This demonstrates that the proposed attentive skip connection is more suitable for depth estimation tasks and that it increases the performance of the network further, in comparison with other attention modules.

In addition, experiments are conducted by using different dilation rates for the attentive skip connection and using various hidden dimensions for the global context module. These experiments are performed to maximize the performance. The results are presented in Table 5. The dilation value of $\{3, 6, 9\}$ provides the best results among those obtained for the various settings. This result supports the notion that applying a well-designed ASPP module for a skip connection can improve the performance of the model by deriving useful features with enlarged receptive fields. With regard to d_h, the value of 512 in the hidden layer provides the best performance among those obtained for the various settings. If the size of the hidden dimension increases, the network usually shows a better performance owing to the increase in depth. However, using an excessively high value for this

TABLE 5. Ablation study on the hyper-parameters. The dilation rate denotes the value of the dilation rate in ASPP for the attentive skip connection. d_h is the size of the intermediate reduced dimension of the global context module. This is validated on the eigen split of the KITTI dataset.

Hyper-Parmas	Value	SqRel	RMSE	SILog
Dilation rate	{3, 6}	0.231	2.694	8.714
	{3, 6, 9}	**0.229**	**2.676**	**8.579**
	{3, 6, 12}	0.234	2.700	8.654
	{3, 6, 12, 18}	0.237	2.731	8.641
d_h	128	0.232	2.697	8.792
	256	0.234	2.684	8.648
	512	**0.229**	**2.676**	**8.579**
	1024	0.242	2.740	8.667

parameter can cause overfitting, and the inference rate will also be adversely affected. Therefore, it is important to find an appropriate value for this task. In summary, based on the ablation study, a d_h value of 512 and dilation rate of {3, 6, 9} are used in this study.

V. CONCLUSION

This paper presents a novel network architecture that leverages the contextual information for monocular depth estimation. Using the proposed modules, the multi-scale attentive skip connections and the global context module, our network captures meaningful contextual representation in the multi-scale and global scale. Extensive experiments and an ablation study demonstrate that the proposed model effectively provides a more accurate predictions, compared to other state-of-the-art methods. Furthermore, our network achieves a significant performance improvement on the KITTI and NYU Depth V2 datasets. Moreover, we plan to investigate the structure of faster and lighter networks to achieve real-time performance.

REFERENCES

[1] A. Bhoi, "Monocular depth estimation: A survey," 2019, *arXiv:1901.09402*. [Online]. Available: http://arxiv.org/abs/1901.09402

[2] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3869–3878.

[3] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[4] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5684–5693.

[5] Y. Gan, X. Xu, W. Sun, and L. Lin, "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 224–239.

[6] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 484–500.

[7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2366–2374.

[8] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 740–756. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46484-8_45

[9] M. Ochs, A. Kretz, and R. Mester, "SDNet: Semantically guided depth estimation network," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2019, pp. 288–302. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-33676-9_20

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[11] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2800–2809.

[12] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

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

[14] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6647–6655.

[15] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3372–3380.

[16] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 283–291.

[17] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1043–1051.

[18] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4106–4115.

[19] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[20] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 53–69.

[21] H. Amini Amirkolaee and H. Arefi, "Monocular depth estimation with geometrical guidance using a multi-level convolutional neural network," *Appl. Soft Comput.*, vol. 84, Nov. 2019, Art. no. 105714.

[22] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C.-F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2624–2632.

[23] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3917–3925.

[24] Y. Chen, H. Zhao, and Z. Hu, "Attention-based context aggregation network for monocular depth estimation," 2019, *arXiv:1901.10137*. [Online]. Available: http://arxiv.org/abs/1901.10137

[25] K. Takagi, S. Ito, N. Kaneko, and K. Sumi, "Boosting monocular depth estimation with channel attention and mutual learning," in *Proc. Joint 8th Int. Conf. Informat., Electron. Vis. (ICIEV) 3rd Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, May 2019, pp. 228–233.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[28] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*. [Online]. Available: http://arxiv.org/abs/1807.06514

[29] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[30] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *Proc. IJCAI*, vol. 7, Jan. 2007, pp. 2197–2203.

[31] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: http://arxiv.org/abs/1405.3531

[32] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "AdaDepth: Unsupervised content congruent adaptation for depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2656–2665.

[33] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[34] Z. Hao, Y. Li, S. You, and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 304–313.

[35] L. Lin, G. Huang, Y. Chen, L. Zhang, and B. He, "Efficient and high-quality monocular depth estimation via gated multi-scale network," *IEEE Access*, vol. 8, pp. 7709–7718, 2020.

[36] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33715-4_54

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "PyTorch: An imperative style, high performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 8024–8035. [Online]. Available: https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

DOYEON KIM (Member, IEEE) received the B.S. degree in computer science from Korea University, Seoul, South Korea, in 2016, and the M.S. degree in robotics program from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018, where she is currently pursuing the Ph.D. degree in electrical engineering. Her research interests include computer vision, deep learning, and machine learning.

SIHAENG LEE (Member, IEEE) received the B.S. degree in mechanical system design engineering from the Seoul National University of Science and Technology, Seoul, South Korea, in 2014, and the M.S. degree from the Division of Future Vehicle, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016, where he is currently pursuing the Ph.D. degree. His research interests include computer vision and deep learning especially with regard to their application to autonomous vehicles.

JANGHYEON LEE (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, deep learning, and machine learning.

JUNMO KIM (Member, IEEE) received the B.S. degree from Seoul National University, Seoul, South Korea, in 1998, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 2000 and 2005, respectively. From 2005 to 2009, he was a Research Staff Member with the Samsung Advanced Institute of Technology (SAIT), South Korea. He joined the Faculty of KAIST, in 2009, where he is currently an Associate Professor of electrical engineering. His research interests include image processing, computer vision, statistical signal processing, machine learning, and information theory.