EdgeConv with Attention Module for Monocular Depth Estimation

Minhyeok Lee Sangwon Hwang Chaewon Park Sangyoun Lee*
Yonsei University
{hydragon516,sangwon1042,chaewon28,syleee}@yonsei.ac.kr

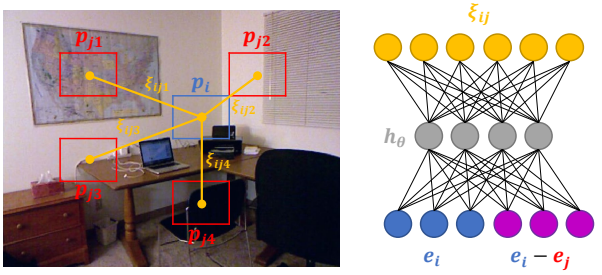
Abstract

Monocular depth estimation is an especially important task in robotics and autonomous driving, where 3D structural information is essential. However, extreme lighting conditions and complex surface objects make it difficult to predict depth in a single image. Therefore, to generate accurate depth maps, it is important for the model to learn structural information about the scene. We propose a novel Patch-Wise EdgeConv Module (PEM) and EdgeConv Attention Module (EAM) to solve the difficulty of monocular depth estimation. The proposed modules extract structural information by learning the relationship between image patches close to each other in space using edge convolution. Our method is evaluated on two popular datasets, the NYU Depth V2 and the KITTI Eigen split, achieving state-of-the-art performance. We prove that the proposed model predicts depth robustly in challenging scenes through various comparative experiments.

1. Introduction

Monocular depth estimation is the task of generating a dense predictive depth map from a single image. The dense depth map provided with an RGB image is useful for understanding the three-dimensional geometric information of the scene in various computer vision tasks. In particular, it is critical to generate accurate depth maps for autonomous driving and robotics, where such information is essential. However, it is difficult to extract 3D structure information from a single monocular image.

To solve this ambiguity, early depth estimation studies [34, 28, 19] used hand-crafted methods based on probabilistic models. Recent approaches based on convolutional neural networks (CNNs) have achieved outstanding performance in numerous computer vision tasks. Owing to the success of deep learning in the computer vision field, many studies have adopted CNNs for monocular depth estimation. The latest studies [43, 31, 48, 22, 6] based on CNNs have demonstrated a better performance improvement than hand-crafted methods. However, these methods do not include a method to extract 3D structural information from a single 2D image. To extract depth information from a single monocular image, it is important to learn the structure information of the image. This structure information can be generated by learning the relationship between adjacent pixels, but few studies use this relationship or extract additional 3D information. Saxena *et al*. [34] introduced the first learning-based depth prediction using the relationship between different points in an image using a Markov random field (MRF). Gan *et al*. [13] proposed the affinity layer, representing the correlation of the values with the surrounding pixels, to extract relative features. Several studies [49, 43, 31] extracted structure information of robust images by generating surface normals as well as depth maps. However, using surface normals requires additional datasets, and the model performance is highly dependent on the surface normal accuracy.



(a) (b)

Figure 1. (a) Visualization of the proposed patch-wise edge convolution. (b) Details of patch-wise edge convolution (PEM). e is the embedded patch p, ξ is the edge features, and h_θ is a nonlinear function with learnable parameter θ.

In this paper, we propose a novel Patch-Wise EdgeConv Module (PEM) and EdgeConv Attention Module (EAM) to address the difficulty of generating structural information from a single image. Unlike the previous attention mechanisms, our proposed method extracts structural information based on the relationship between image patches using edge convolution. Edge convolution (EdgeConv) proposed by [41] is based on graph convolution and learns structural information by extracting edge feature information between

critical nodes of the 3D point cloud datasets. As shown in Figure 1, our PEM extracts edge features based on the relationship between adjacent patches in the feature space. To the best of our knowledge, our model is the first method to employ EdgeConv for depth estimation. The EAM generates a patch-wise attention map using edge feature information extracted by the PEM from multi-scale feature maps. To reconstruct 3D structure information from a single 2D image, it is essential to learn edge features representing the connection structure between adjacent pixels. The EAM combines the local structure information generated by the PEM and the global feature to make the model more robust to monocular depth estimation.

We conduct comprehensive evaluations on two challenging datasets, KITTI Eigen split [15] and NYU Depth V2 [35]. The results demonstrate that the proposed method achieves significant improvements over state-of-the-art methods, both qualitatively and quantitatively. Specifically, our method achieves the lowest root mean square error (RMSE) on both datasets.

Our main contributions can be summarized as follows:

- We propose the novel PEM and EAM, which learns structural information between image patches by applying edge convolution.
- Our modules significantly improve the monocular depth estimation performance in challenging scenes by extracting edge features between patches close in feature space and learning the relationship between patches.
- The proposed network achieves state-of-the-art performance on the NYU Depth V2 [35] and KITTI [15] datasets. In addition, we prove the effectiveness of the proposed module through various comparative experiments.

2. Related Work

Supervised Monocular Depth Estimation. Monocular depth estimation has been studied extensively. Saxena *et al*. [34] used discriminatively trained MRF to obtain a functional mapping of depth from visual cues. Liu *et al*. [26] used semantic segmentation and depth reconstruction to extract context information for depth estimation. Eigen *et al*. [9] proposed a multi-scale convolutional architecture that predicted the coarse global depth and refined it progressively. Xu *et al*. [45] applied Conditional Random Field to the model and used attention mechanisms to learn multi-scale features. Xu *et al*. [43] proposed four multi-task learning methods for depth estimation: monocular depth prediction, surface normal estimation, semantic segmentation, and contour detection. Their paper argued that it could provide multi-modal information through multi-task learning. Qi *et al*. [31] proposed a mapping model between depth and surface normals to extract geometric information of images. Zhang *et al*. [48] introduced a task-recursive learning model that repeatedly performs depth estimation and semantic segmentation tasks. They also designed a task-attentional module to enhance the connectivity between the two tasks. Lee *et al*. [22] applied local planar guidance layers to the decoder to generate high-quality depth maps. Wang *et al*. [39] performed hierarchical embedding by applying depth-based scene classification and depth reconstruction. Chen *et al*. [6] improved estimation performance by proposing a pixel-level attention model that effectively captured the pixel-level and image-level context.

Attention Mechanism. Attention mechanisms have provided significant improvements in machine translation and natural language processing. They have been used recently in various computer vision fields. Non-local neural networks [40] learn the relationships between pixels in different locations in the image. Fu *et al*. [12] improved semantic segmentation performance using two types of non-local blocks. Zhang *et al*. [47] introduced a better image generator with non-local operations. In addition, attention mechanisms can emphasize important spatial information of feature maps. For example, [30, 42] meaningfully improved the classification task performance by adding channel attention and spatial attention mechanisms to the model.

Graph Convolutional Networks. Unlike CNNs, which learn features between image pixels, graph convolutional neural networks learn important features from connected nodes of the graph. Graph convolutional neural networks collect information from nodes and edges within the graph by applying convolution operations to the graph. Defferrard *et al*. [7] introduced convolution for a graph as a multiplication for its spectral representation. Some studies [29, 36, 1] focused on the convolution operation on the surface represented by the graph. In particular, graph convolution is often used to effectively obtain 3D structure information from point cloud data consisting of a large number of points. Representatively, DGCNN [41] proposed EdgeConv, a type of graph convolution, to learn semantic displacement between key points and feature space neighbors.

3. Proposed Approach

3.1. Overall Architecture

The overall architecture is shown in Figure 2. Our model is based on the feature pyramid network (FPN) architecture [25]. It includes two of the proposed PEMs and one EAM. The PEMs extract multi-scale edge features between patches of feature maps from the model encoder layers at different locations. The EAM performs patch-wise attention using the global depth contextual information of the

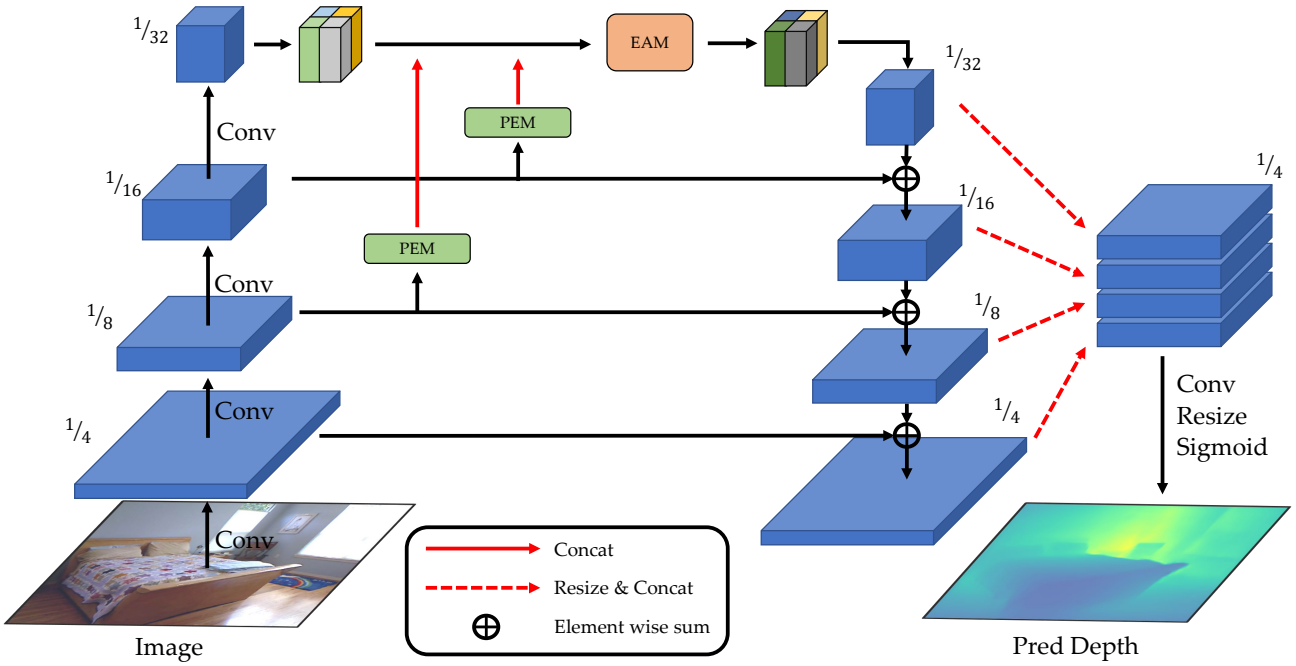


Figure 2. Proposed architecture is an end-to-end framework based on the feature pyramid network (FPN) [25]. The proposed PEM is located in the skip connection of the feature map layers of different sizes. The proposed EAM is located after the last layer of the encoder.

last layer of the encoder and the multi-scale local edge feature. Finally, we resize and stack the decoder's multi-scale feature maps and pass them through the convolution and sigmoid layers. We multiply the output with the predefined maximum depth to get real depth values. In the following sections, we will cover the PEM and EAM in detail.

3.2. Patch-Wise EdgeConv Module (PEM)

The PEM extracts patch-wise edge features corresponding to local structure information using EdgeConv [41]. EdgeConv is designed for point cloud datasets, so it cannot be directly applied to 2D images.

To solve this problem, the first step of the PEM is to divide the input feature maps into patches and generate linear embedding sequences of them. This process is the patch embedding proposed by [8]. The patch embedding function f_e is expressed as $f_e : p_i \in \mathbb{R}^{c_f \times w_p \times h_p} \rightarrow e_i \in \mathbb{R}^{c_e}$, where (w_p, h_p) is the patch size, c_f is the number of channels of the feature map after the 1×1 convolution layer, c_e is the length of the embedded vector, p_i is the patch, and e_i is the embedded patch at position i. Therefore, the size of the patch-embedded input feature map X_e is $c_e \times N_p$, where N_p is the number of patches. We set $N_p = \frac{H}{32} \times \frac{W}{32}$, where the height of the input image is H and the width is W. The detailed reasons are covered in section 3.3.

The second step of the PEM is to apply the EdgeConv Module (EM) as shown in Figure 3 (a). The EM extracts edge features after generating graphs between patches using the k-Nearest-Neighbor (k-NN) algorithm. As shown in Figure 1, we define edge features as $\xi_{ij} = h_\theta\left(e_i, e_j - e_i\right)$, where $h_\theta : \mathbb{R}^{c_e} \times \mathbb{R}^{c_e} \rightarrow \mathbb{R}^{c_o}$ is a nonlinear function with a set of learnable parameters θ and c_o is the length of the output edge feature ξ_{ij}. This process is equivalent to generating dynamic graphs proposed by [41], and the output of this process is a tensor in the shape of $a_k \times c_o \times N_p$, where a_k is the number of nearest neighbors.

In the last step, the local edge feature map $\xi_o \in \mathbb{R}^{c_o \times N_p}$ is obtained by passing it through the channel-wise multi-layer perceptron (MLP) layer. As shown in Figure 2, we extract multi-scale edge features by applying the PEMs to two feature maps of different resolutions. The two edge feature maps are used as inputs to the EAM.

3.3. EdgeConv Attention Module (EAM)

The EAM performs patch-wise attention from edge features generated from the PEM using a self-attention mechanism. The structure of the proposed EAM is shown in Figure 3 (b). The EAM input is defined as follows:

$$X_{cat} = \text{concat}\left(f_g; \xi_{o/8}; \xi_{o/16}\right) \tag{1}$$

When the height of the input image is H and the width is W, $f_g \in \mathbb{R}^{c_g \times \left(\frac{W}{32} \times \frac{H}{32}\right)}$ is the spatially flattened feature map

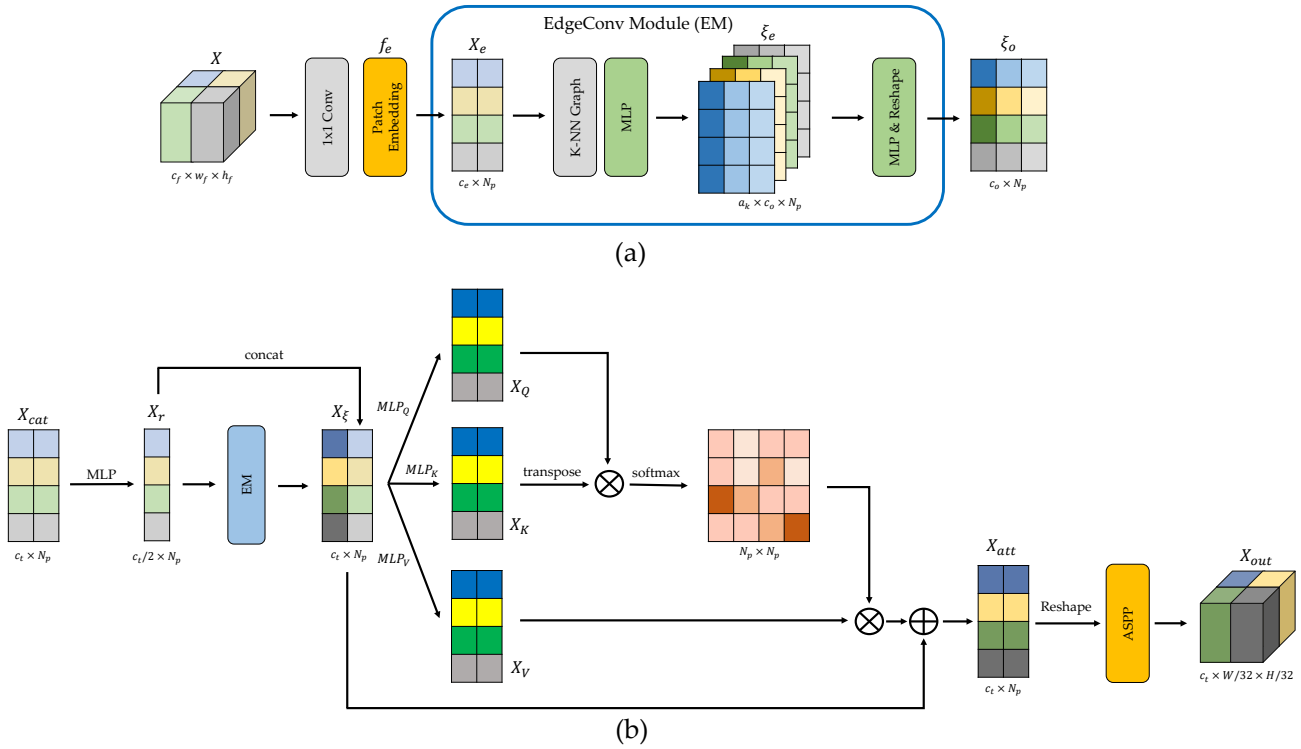


Figure 3. (a) Structure of Patch-Wise EdgeConv Module (PEM). (b) Structure of EdgeConv Attention Module (EAM).

of the last layer of the encoder, where c_g is the number of channels. $\xi_{o/8} \in \mathbb{R}^{c_o/8 \times N_p}$ and $\xi_{o/16} \in \mathbb{R}^{c_o/16 \times N_p}$ are the results of the PEMs for encoder feature maps with $1/8$ and $1/16$ size input images. Because we set $N_p = \frac{H}{32} \times \frac{W}{32}$, the size of X_{cat} is $c_t \times N_p$, where $c_t = c_g + c_{o/8} + c_{o/16}$.

The next step is to extract the edge features using the EM defined in section 3.2. This process is defined as follows:

$$X_\xi = \text{concat}\left(X_r; \text{EM}\left(X_r\right)\right) \tag{2}$$

X_{cat} passes through the channel-wise MLP layer to become $X_r \in \mathbb{R}^{\frac{c_t}{2} \times N_p}$, which has half the number of channels of X_{cat}. Because we set the output channel of EM to $\frac{c_t}{2}$ as shown in Figure 3 (b), the size of X_ξ is the same as X_{cat}.

The final process of the EAM is the MLP blocks and self-attention module (SAM). The self-attention module input is a triplet of X_K (key), X_Q (query), and X_V (value), similar to [38]. The three inputs of the SAM are expressed as $X_K = \text{MLP}_K\left(X_\xi\right), X_Q = \text{MLP}_Q\left(X_\xi\right)$ and $X_V = \text{MLP}_V\left(X_\xi\right)$. The self-attention formula is:

$$\text{SAM}\left(X_Q, X_K, X_V\right) = \text{softmax}\left(\frac{X_Q X_K^T}{\sqrt{d}}\right) \cdot X_V, \tag{3}$$

where d is the channel of the SAM inputs. The final EAM output $X_{att} \in \mathbb{R}^{c_t \times N_p}$ is expressed as follows:

$$X_{att} = X_\xi + \text{SAM}\left(X_Q, X_K, X_V\right) \tag{4}$$

In addition, X_{att} is reshaped to $c_t \times \frac{W}{32} \times \frac{H}{32}$. Finally, we connect the ASPP [3] layer to extract global contextual information efficiently by increasing the receptive field of the model. We then feed the output of the ASPP layer X_{out} to the decoder.

3.4. Objective Functions

We apply the scale-invariant error proposed by Eigen *et al.* [10] as the training loss. The objective function is defined as follows:

$$L = \sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}\left(\sum_i g_i\right)^2}, \tag{5}$$

where $g_i = \log d_i - \log d_i^*$ with the ground truth depth d_i and the predicted depth d_i^*. T is the number of valid pixels of the ground truth. Specifically, the loss is a mixture of elementwise l_2 and the scale-invariant error. Because it is the sum of the variance and the weighted squared mean of the error in log space, setting a high λ enforces a greater focus on minimizing the variance of the error. Therefore, we set λ to 0.85, the same as in [22].

Method	cap	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	SqRel $\downarrow$	RMSE $\downarrow$	RMSE *log* $\downarrow$
Eigen et al. [10]	0-80m	0.702	0.898	0.967	0.203	1.548	6.307	0.282
Godard et al. [16]	0-80m	0.861	0.949	0.976	0.114	0.898	4.935	0.206
Kuznietsov et al. [21]	0-80m	0.862	0.960	0.986	0.113	0.741	4.621	0.189
Gan et al. [13]	0-80m	0.890	0.964	0.985	0.098	0.666	3.933	0.173
Fu et al. [11]	0-80m	0.932	0.984	0.994	0.072	0.307	2.727	0.120
Yin et al. [46]	0-80m	0.938	0.990	0.998	0.072	-	3.258	0.117
Lee et al. [22]	0-80m	0.956	0.993	0.998	**0.059**	0.245	2.756	0.096
⋆ MobileNetV2 [33]	0-80m	0.935	0.991	0.998	0.071	0.291	2.970	0.110
⋆ Densenet161 [18]	0-80m	0.944	0.992	0.998	0.069	0.272	2.829	0.099
⋆ EfficientNet-B6 [37]	0-80m	0.950	0.992	0.998	0.067	0.254	2.785	0.098
Ours-MobileNetV2	0-80m	0.947	0.992	0.998	0.065	0.261	2.925	0.107
Ours-Densenet161	0-80m	0.956	**0.994**	**0.999**	0.063	0.243	2.711	0.096
Ours-EfficientNet-B6	0-80m	**0.958**	**0.994**	**0.999**	0.060	**0.231**	**2.642**	**0.094**
Garg et al. [14]	0-50m	0.740	0.904	0.962	0.169	1.080	5.104	0.273
Godard et al. [16]	0-50m	0.873	0.954	0.979	0.108	0.657	3.729	0.194
Kuznietsov et al. [21]	0-50m	0.875	0.964	0.988	0.108	0.595	3.518	0.179
Gan et al. [13]	0-50m	0.898	0.967	0.986	0.094	0.552	3.133	0.165
Fu et al. [11]	0-50m	0.936	0.985	0.995	0.071	0.268	2.271	0.116
Lee et al. [22]	0-50m	**0.964**	0.994	**0.999**	**0.056**	0.169	1.925	**0.087**
⋆ MobileNetV2 [33]	0-50m	0.943	0.992	0.998	0.070	0.253	2.539	0.101
⋆ Densenet161 [18]	0-50m	0.950	0.993	0.998	0.067	0.202	2.014	0.091
⋆ EfficientNet-B6 [37]	0-50m	0.958	0.993	0.998	0.061	0.182	1.993	0.089
Ours-MobileNetV2	0-50m	0.958	0.993	0.998	0.068	0.221	2.127	0.098
Ours-Densenet161	0-50m	0.963	0.994	**0.999**	0.060	0.177	1.960	0.088
Ours-EfficientNet-B6	0-50m	**0.964**	**0.995**	**0.999**	0.057	**0.168**	**1.897**	**0.087**

Table 1. Performance comparison with other state-of-the-art methods on the KITTI Eigen split. $\uparrow$ indicates that higher is better and $\downarrow$ indicates that lower is better. Baseline models are marked with $\star$ symbol. The top two results are marked with **bold** and underline. The same notation is used in the following sections. All methods were evaluated on the split given by Eigen *et al*. [10]. Our approach achieved state-of-the-art results.

Method	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$	log10 $\downarrow$
Eigen et al. [9]	0.769	0.950	0.988	0.158	0.641	-
Chakrabarti et al. [2]	0.806	0.958	0.987	0.149	0.620	-
Li et al. [24]	0.789	0.955	0.988	0.152	0.611	0.064
Xu et al. [44]	0.811	0.954	0.987	0.121	0.586	0.052
Lee at al. [23]	0.815	0.963	0.991	0.139	0.572	-
Qi et al. [31]	0.834	0.960	0.990	0.128	0.569	0.057
Hu et al. [17]	0.866	0.975	0.993	0.115	0.530	0.050
Chen et al. [5]	0.878	0.977	0.994	0.111	0.514	0.048
Yin et al. [46]	0.875	0.976	0.994	0.108	0.416	0.048
Lee et al. [22]	0.885	0.978	0.994	0.110	0.392	0.047
Chen et al. [4]	**0.899**	0.983	0.996	**0.098**	0.376	**0.042**
⋆ MobileNetV2 [33]	0.805	0.962	0.991	0.149	0.488	0.062
⋆ Densenet161 [18]	0.862	0.975	0.994	0.122	0.416	0.051
⋆ EfficientNet-B6 [37]	0.871	0.981	0.996	0.115	0.399	0.049
Ours-MobileNetV2	0.838	0.969	0.993	0.135	0.439	0.055
Ours-Densenet161	0.881	0.981	0.996	0.113	0.387	0.047
Ours-EfficientNet-B6	0.893	**0.985**	**0.997**	0.107	**0.373**	0.046

Table 2. Performance comparison with other state-of-the-art methods on the NYU Depth V2 test set.

4. Experiments

4.1. Datasets and Evaluation Metrics

NYU Depth V2 Dataset. The NYU Depth V2 dataset [35] contains 464 indoor scenes, consisting of 120K images and paired depth maps with a resolution of 640×480 pixels. We followed the official train/test split from previous studies, using 249 scenes for training and 215 scenes (654 images) for testing. Due to the asynchronous capture rates between the original RGB images and the depth maps, all the images and depth maps used for the experiments on NYU Depth V2 were collected from Lee *et al*. [22]. In the evaluation, we applied a predefined center cropping method by Eigen *et al*. [10].

KITTI Dataset. The KITTI dataset [15] contains scenes from 61 outdoor driving scenarios captured by multiple sensors. The dataset contains RGB images with a resolution of 1241×375 pixels and the corresponding LiDAR point clouds. We used the official KITTI dataset containing RGB images and post-processed depth maps of projected LiDAR point clouds as the ground truth. The ground truth is produced by combining LiDAR scans. In the experiments, we followed the common data split proposed by Eigen *et al*. [10] for comparison with previous studies. Therefore, 697 images covering a total of 29 scenes were used for evaluation, and the remaining 23,488 images covering 32 scenes were used for training. For the KITTI dataset, we trained our networks on 1216×352 pixels using a bottom-center crop.

Evaluation Metrics. Following the previous work [10], we

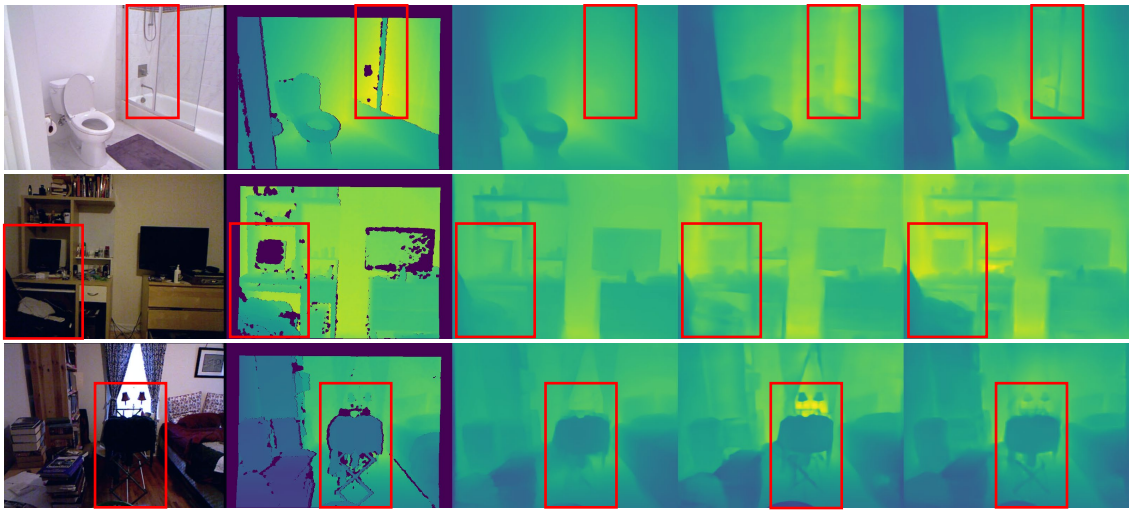


Figure 4. Comparison of qualitative results for the NYU Depth V2 test set with other state-of-the-art methods by Yin *et al*. [46] and Lee *et al*. [22]. Our method is based on the EfficientNet-B6 encoder.

evaluate our methods using the following quantitative metrics:

- Accuracy with threshold: percentage (%) of d_i s.t. $max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = \delta < thr, thr = 1.25, 1.25^2, 1.25^3$
- RMSE : $\sqrt{\frac{1}{N}\sum_i^N (d_i - d_i^*)^2}$
- RMSE *log* : $\sqrt{\frac{1}{N}\sum_i^N ||\log_{10} d_i - \log_{10} d_i^*||^2}$
- AbsRel: $\frac{1}{N}\sum_i^N \frac{|d_i - d_i^*|}{d_i^*}$
- SqRel: $\frac{1}{N}\sum_i^N \frac{||d_i - d_i^*||^2}{d_i^*}$

, where d_i and d_i^* is the ground-truth depth and the estimated depth at pixel i, respectively; N denotes a collection of pixels that the ground truth values are available. In addition, we used mean log10 error (log10) for comparison with [22].

4.2. Implementation Details

We implemented the proposed method using the open deep learning framework PyTorch. For network training, we used the Adam optimizer [20] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$. The learning rate decayed from 10^{-4} to 10^{-5} with the cosine annealing scheduler [27]. The total number of epochs was set to 50 with batch size 4 with a single NVIDIA Quadro RTX 6000 GPU for all experiments in this study.

For the NYU Depth V2 dataset, we set the patch sizes of the patch embedding functions f_e in the two PEMs to 4×4 and 2×2. Because each input feature map resolution was 80×60 and 40×30, the number of patches N_p generated by each patch embedding function was 300. For the KITTI dataset, the size of each PEM's input feature map was 152×44 and 76×22. Because we applied the same patch size to the KITTI dataset, the number of patches was 418.

We used EfficientNet-B6 [37], DenseNet161 [18], and MobileNetV2 [33], which were pretrained on image classification using the ImageNet-1K dataset [32], as the backbone networks. To avoid overfitting, we applied data augmentation to all the experiments. Random horizontal flipping was applied in all cases. We also randomly rotated the input images in ranges of $[-1, 1]$ and $[-2.5, 2.5]$ degrees for the KITTI and NYU Depth V2 datasets, respectively. Thereafter, contrast, brightness, and color adjustment were applied to the inputs for both datasets.

4.3. Results

Tables 1 and 2 compare the performance results of the proposed method with previously presented state-of-the-art algorithms for the NYU Depth V2 and KITTI datasets. The proposed method was evaluated with the backbone encoders MobileNetV2 [33], DenseNet161 [18], and EfficientNet-B6 [37]. All these models used PEM and EAM, as shown in Figure 2, and each performance effect is described in detail in the ablation study section.

NYU Depth V2 Dataset. As shown in Table 2, our

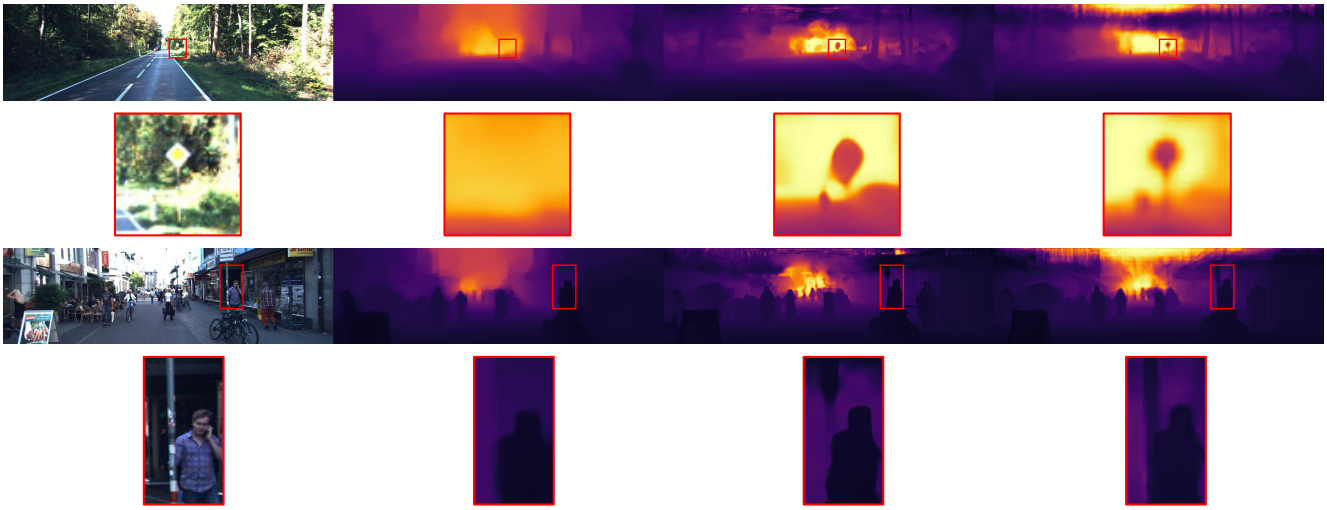


Figure 5. Comparison of qualitative results for the KITTI Eigen split with other state-of-the-art methods by Yin *et al.* [46] and Lee *et al.* [22]. Our method is based on the EfficientNet-B6 encoder.

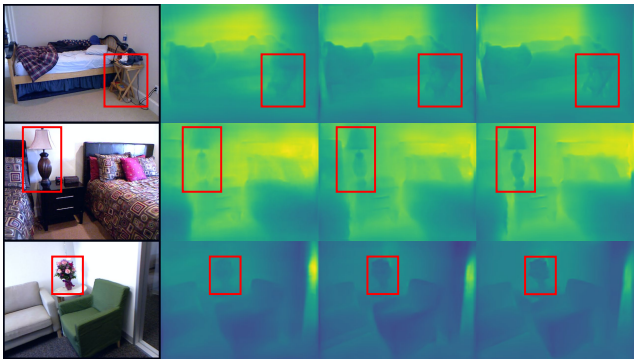


Figure 6. Comparison of qualitative results for the NYU Depth V2 test set using the baseline model (A), baseline with PEM (B), and the proposed PEM and EAM combination (D). A, B, and D are the model names shown in Table 3. The baseline model is MobileNetV2.

EfficientNet-B6 based model outperforms the previous methods [9, 2, 24, 44, 23, 31, 17, 5, 46, 22, 4] in the NYU Depth V2 dataset. In particular, the RMSE was the lowest, at 0.373. RMSE measures depth errors directly and applies a heavier penalty for undesirable larger errors. Therefore, RMSE is known as the main evaluation metric of depth estimation.

The NYU Depth V2 dataset is an indoor dataset containing a large amount of furniture with complex surface structures. Moreover, light reflections and shadows caused by windows and glass make it difficult to predict the depth map. Therefore, to predict the depth of areas lacking visual cues, it is important for the model to extract structural information from the relationships between adjacent pixels.

Figure 4 shows the qualitative results of our method for various challenging indoor scenes. The results in Figure 4 show that our method robustly predicts depth in extreme lighting conditions. In particular, our model generates accurate depth maps in low-light environments such as the second row and intense lighting conditions such as the third row of Figure 4. Furthermore, the first row of Figure 4 shows the results of our method in an optical illusion situation caused by an obstacle on the glass. However, the proposed PEM and EAM help the model reinforce the learning of relationships between neighboring patches to make accurate depth prediction possible.

KITTI dataset. In general, outdoor scenes are more challenging because the lighting changes are more severe than in indoor scenes. The KITTI dataset contains images of outdoor driving scenarios for autonomous vehicles. These images include many trees and buildings with complex structures, along with traffic signs with thin columns, making it difficult to predict the depth. Table 1 shows the performance of the proposed model for the KITTI dataset compared to the previous methods [10, 16, 21, 13, 11, 46, 22, 14]. As shown in Table 1, our EfficientNet-B6-based model achieved state-of-the-art performance in most evaluation metrics. The qualitative results in Figure 5 support the standpoint that our model also robustly estimated depth for outdoor scenes. In particular, the proposed method accurately predicted depth information for long and thin objects such as traffic signs and traffic cones. Furthermore, our model produced results that were more robust to changes in brightness. This is because the PEM and EAM extract structural information from the relationships between patches, and thus they supplement the lack of visual cues with related adjacent patches. We detail the effects of the proposed

name	Baseline	EAM	PEM	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	RMSE $\downarrow$	log10 $\downarrow$
A	✓			0.805	0.962	0.991	0.149	0.488	0.062
B	✓		✓	0.823	0.964	0.992	0.140	0.461	0.060
C	✓	✓		0.835	0.968	0.992	0.136	0.443	0.056
D	✓	✓	✓	**0.838**	**0.969**	**0.993**	**0.135**	**0.439**	**0.055**

Table 3. Evaluation results of the baseline model (MobileNetV2) and the proposed model with the PEM and EAM combination for the NYU test set.

name	Baseline	EAM	PEM	cap	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	AbsRel $\downarrow$	SqRel $\downarrow$	RMSE $\downarrow$	RMSE *log* $\downarrow$
A	✓			0-80m	0.935	0.991	**0.998**	0.071	0.291	2.970	0.110
B	✓		✓	0-80m	0.939	0.991	**0.998**	0.068	0.273	2.956	0.109
C	✓	✓		0-80m	0.946	**0.992**	**0.998**	**0.065**	0.269	2.931	**0.107**
D	✓	✓	✓	0-80m	**0.947**	**0.992**	**0.998**	**0.065**	**0.261**	**2.925**	**0.107**

Table 4. Evaluation results of the baseline model (MobileNetV2) and the proposed model with the PEM and EAM combination for the KITTI Eigen split.

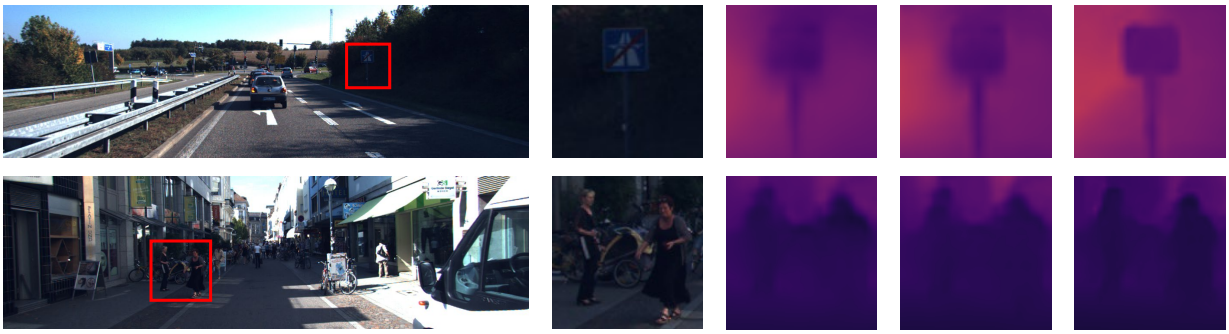


Figure 7. Comparison of qualitative results of the baseline model (MobileNetV2), baseline with PEM, and the proposed model with the PEM and EAM combination for the KITTI Eigen split. A, B, and D are the model names shown in Table 4.

PEM and EAM in the next section.

Impact of the PEM and EAM. We performed an ablation study on the NYU Depth V2 and KITTI datasets to further demonstrate the impact of the proposed PEM and EAM. The baseline model encoder for the ablation study was MobileNetV2, and the implementation details of all the experiments were the same as in Section 4.2. Tables 3 and 4 show our experimental results for the NYU Depth V2 and KITTI datasets, respectively. As shown in Tables 3 and 4, the model has the highest depth estimation performance when the PEM and the EAM were all used. The EAM, including EdgeConv, learns structure information by extracting edge features between patches, like the PEM. However, using the PEM, the model can learn more diverse inter-patch relationships from multi-scale feature maps. Therefore, the performance was further improved with the PEM. Figures 6 and 7 show the qualitative results of the ablation studies. Our model more accurately estimated the depth map and generated sharp object contours than the baseline model. This is because both global contextual information and structural information are important in monocular depth estimation. The experimental results support that the proposed model learned structural information by learning the relationships between nearby pixels.

5. Conclusion

In this paper, we propose the novel PEM and EAM for monocular depth prediction. We apply EdgeConv for depth prediction to extract edge features between patches close to each other in the feature space. The proposed method makes robust depth prediction by learning structural information through edge features. Our model achieves state-of-the-art performance on NYU Depth V2 and KITTI datasets, especially with the smallest RMSE. The ablation study proves the effectiveness of the proposed modules and shows that the depth prediction performance is greatly improved in challenging scenes.

Acknowledgement. This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00172, The development of human Re-identification and masked face recognition based on CCTV camera)

References

[1] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[2] Ayan Chakrabarti, Jingyu Shao, and Gregory Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. *arXiv preprint arXiv:1605.07081*, 2016.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[4] Tian Chen, Shijie An, Yuan Zhang, Chongyang Ma, Huayan Wang, Xiaoyan Guo, and Wen Zheng. Improving monocular depth estimation by leveraging structural awareness and complementary datasets. *arXiv preprint arXiv:2007.11256*, 2020.

[5] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. *arXiv preprint arXiv:1907.06023*, 2019.

[6] Yuru Chen, Haitao Zhao, Zhengwei Hu, and Jingchao Peng. Attention-based context aggregation network for monocular depth estimation. *International Journal of Machine Learning and Cybernetics*, pages 1–14, 2021.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375*, 2016.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[13] Yukang Gan, Xiangyu Xu, Wenxiu Sun, and Liang Lin. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 224–239, 2018.

[14] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[17] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[19] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017.

[22] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[23] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 330–339, 2018.

[24] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017.

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[26] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260. IEEE, 2010.

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[28] Aamir Saeed Malik and Tae-Sun Choi. A novel algorithm for estimation of depth map using image focus for 3d shape recovery in the presence of noise. *Pattern Recognition*, 41(7):2200–2225, 2008.

[29] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

[30] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[31] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[34] Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. Learning depth from single monocular images. In *NIPS*, volume 18, pages 1–8, 2005.

[35] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[36] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017.

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[39] Lijun Wang, Jianming Zhang, Yifan Wang, Huchuan Lu, and Xiang Ruan. Cliffnet for monocular depth estimation with hierarchical embedding loss. In *European Conference on Computer Vision*, pages 316–331. Springer, 2020.

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[41] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[43] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[44] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.

[45] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3917–3925, 2018.

[46] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019.

[47] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[48] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018.

[49] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019.