Multi-Scale Context Enhanced Network for Monocular Depth Estimation

Wenju Wang1, Yue Ning1*

1College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China

*Yue Ning (e-mail: 183732492@st.usst.edu.cn)

Abstract. Monocular depth estimation is a classical computer vision task. At present, most CNN methods cannot effectively combine high-level and low-level features, leading to the loss of details and blurring of boundaries. To solve the problem, we propose a Multi-Scale Context Enhanced Network (MCEN) to learn more abundant context and expand its receptive field for high-accuracy estimation. Our method employs CRE-HRNet (Context and Receptive Enhanced High-Resolution Network) with four branches ranging from low-dimension to high-dimension features to obtain richer contextual information and extract multi-scale features. It then uses RM (Refinement Module) adopting the residual dilated convolution to retain detailed information and improve the receptive field. Finally, non-local block enables our network to capture the long-distance context through its special non-local operation. Experiments with the NYU Depth V2 dataset show its outstanding performance.

1. Introduction

Depth estimation is a classical computer vision task. An accurate depth map contributes to scene reconstruction, robot navigation, semantic segmentation, etc. However, a monocular image often presents difficult physical scenes leading to fuzzy uncertainty within the monocular depth estimation problem [1], complicating a high-accuracy depth estimation.

To overcome inherent ambiguities, LiDARs and depth sensors (such as RGB-D camera) are widely used to obtain geometric information. But both the method have their own limitations. For example, LiDARs generate the sparse cloud with its cost-prohibitive equipment. Geometry-based methods estimate depth by solving and restoring a 3D structure from serialized images using geometric constraints. SFM (Structure from Motion) [2] is the most representative approach, determining the depth of sparse features through feature correspondences and geometric constraints within image sequences. But the method is easily defeated by the external environment, leading to uncertainty in constructing high-accuracy depth maps. How to get the accurate depth map from a monocular image is still a challenge lacking of geometry constrains.

With the rapid development of deep learning [3], deep neural networks show their remarkable performance in image processing, such as pattern recognition, semantic segmentation and scene understanding. For this reason, researchers have applied it to monocular depth estimation. Eigen et al. [1] proposed a multiple scale depth network (MSDN) and a dimensional constant-square loss function in a CNN, enabling monocular depth estimation from coarse to fine. To further improve accuracy, Eigen et al. [4] added a new upper sampling sub-functional network built on MSDN and applied it to semantic segmentation, surface vector prediction, and other tasks. Combining with CNN, Li et al. [5] proposed a super-resolution depth map reconstruction algorithm able to extract feature information autonomously.
to simulate the human visual system. However, with the same number of samples, overfitting is still a challenge with more layers. Xie et al. [6] used a skip-connection method to merge a low-resolution high-rise network with a high-resolution low-level network to improve the continuity between images with different resolutions, but the model involved more parameters affecting the training efficiency. CNNs improve the receptive field in monocular depth estimation, capturing long-range dependencies between residual unit by backpropagation. Local operations, however, are difficult to implement and prone to feature loss in the middle tier, which results in estimates with low accuracy.

To construct the high-accuracy depth maps, we propose a monocular depth estimation framework called MCEN (Multi-scale Context Enhanced Network). In general, the contributions of this paper can be listed as follows:

1) We propose CRE-HRNet obtaining richer contextual information and extracting multi-scale features from monocular image to achieve a higher accuracy depth map.
2) We propose a residual dilated convolution module called RM (Refinement Module) to reduce the detail loss and produce a more abundant depth image.
3) Experimental results demonstrate that our proposed method achieves state-of-the-art performance in quantitative metrics on the public NYU Depth V2 dataset.

2. Methodology

In this section, we present our Multi-Scale Context Enhanced Network (MCEN) for monocular depth estimation. As shown in Figure 1, our framework is composed of three modules: CRE-HRNet, RM, and non-local block.

2.1. Network Architecture

First, we input monocular RGB image of size H×W×C. After a downsampling operation, the feature maps of size H/2×W/2×2C are used as the MCEN input. Following CRE-HRNet (Context and Receptive Enhanced High-Resolution Network) uses four branches ranging from low-dimension to high-dimension features to extract multiple scale features extractor for the input monocular RGB image. The RM (Refinement Module) adopts the residual dilated convolution to produce a more abundant depth image. Non-local block enables the network to capture distant pixels throughout the scenes through its special non-local operation. We describe the details in the following sections.

2.2. CRE-HRNet

To extract feature information from monocular RGB image and obtain richer contextual information, we incorporate CRE-HRNet to our depth estimation. We incorporate HRNET-W48 [7] in our CRE-HRNet method. It has two sub-modules, as shown in Figure 1. (1). The feature pyramid (abbr. as FP...
hereafter) is a four-layer feature pyramid structure with resolution descending from top to bottom. The multiple scale high resolution module is a four-branch high-resolution network. These two submodules perform multi-scale feature extraction as detailed below.

To extract the shallow-to-deep depth feature information from monocular image, we use the FP, as shown in Figure 1. (1-a). As noted previously, HRNET-W48 is the backbone of CRE-HRNet, with W48 representing the number of dimensions of the first layer of the FP, 4C=48. Downsampling the feature map in each layer of the FP reduces the width by half and doubles the number of dimensions. Thus, our four-layer pyramid structure incrementally produces feature maps of size $H\times W/4\times 4C$, $H/8\times W/8\times 8C$ and $H/16\times W/16\times 16C$. The feature map $H/2\times W/2\times 2C$ is used as the input to the FP. Features such as those extracted by the feature pyramids have been widely used in semantic segmentation [8] and object detection [9]. Combined with the extraction of features at different scales, this approach solves the problem of missing details in the lower sampling process and improves its accuracy.

The multi-scale high-resolution module consists of four branches with resolution decreasing from top to bottom, named Stages 1, 2, 3, and 4. Stage 1 is located in the high-resolution main branch, as shown in the gray area of Figure 1. (1-b). It continually receives characteristic information from the lower resolution branches Stage 2, Stage 3, and Stage 4, and fuses context and depth information to improve the depth estimation accuracy.

As the above two modules work together, the specific implementation process is as follows. The FP connects the feature maps $H/4\times W/4\times 4C$, $H/8\times W/8\times 8C$ and $H/16\times W/16\times 16C$ together horizontally. These four feature maps with different scales are used as the input of the multi-scale high-resolution module, which fuses the multiple level features using element-wise addition:

$$F_k = P_k + S_k$$  \tag{1}

where $P_k$ represents the feature mapping of the kth level of the FP, $S_k$ represents the kth branch ($k\geq 1$, $k=2,3,4$) of the CRE-HRNet, and $F_k$ represents the output of the feature map between the two after fusion.

In particular, the first-level branch represents the main high-resolution branch, as shown in the gray area of Figure 1. (1-b). The first-level feature map of Feature Pyramid is directly used as the input of the high-resolution main branch. As shown in Figure 1. (1) after four levels of feature extraction and fusion, it will contain four different scales of feature information from the FP. The feature maps can be used as the input of the following the Refinement Module (abbr. as RM hereafter).

2.3. Refinement Module

Dilated convolution has a strong ability to mine image clues and can increase the receptive field without reducing the resolution. It also effectively learns global context feature information to achieve good results in semantic segmentation [10]. Depth estimation is similar to semantic segmentation because they both take RGB images or grayscale images as their input and then estimate their pixels. Our proposed RM outputs a high resolution feature map used to find image clues and estimate the depth of various objects accurately from complex indoor scenes. Its core unit is dilated convolution. A smaller dilation rate enables the kernel to learn more detailed information. A larger dilation rate produces a larger receptive field but with the detail loss. Our RM adopts a residual dilated convolution, applying larger dilation rates after smaller ones. This method retains detailed information from the complex scene while improving its receptive field. RM consists of three branches as shown in Figure 1. (2). We define the structure of the different dilation rates within each branch as extended residual units denoted $I_1$, $I_2$, and $I_3$, respectively. RM consists of three branches denotes $y_1$, $y_2$, and $y_3$.

The first branch consists of an expansion residual unit, with dilation rates of 3, 6, and 12:

$$y_1 = \phi_2(\phi_3(F_k))$$  \tag{2}
As shown in (2), \( \varphi_x(F_k) \) represents the output value of the residual cell with the dilation rate of 3. \( F_k \) represents the feature map from the CRE-HRNet output, and \( y_i \) represents the output value of the first branch.

The second branch contains expansion residual units with dilation rates of 3 and 6:

\[
y_2 = \varphi_k(\varphi_x(F_k))
\]

(3)

The third branch consists of expansion residual units with dilation rate 3:

\[
y_3 = \varphi_3(F_k)
\]

(4)

As shown in Figure 1. (2), \( F_k \) represents the final connection trunk belonging to the unit map. RM also performs element-wise addition to fuse the feature information of the four parallel branches. Accordingly, the output of the RM can be defined as:

\[
y = \varphi_2(\varphi_x(F_k)) + \varphi_1(\varphi_x(F_k)) + \varphi_3(F_k) + F_k
\]

(5)

As mentioned above, RM implements stacking and parallel cascading. The expansion residual unit of the previous branch accepts the input of the expansion residual unit of the neighboring or cross-level branches. This effectively improves the receptive field and accepts more abundant context information without losing spatial information. Because the outputs of the three branches of the RM are fused, its output embeds different contextual feature information in a larger range. Therefore, our proposed RM has the advantage of expanding residual cascades.

2.4. Non-local block

The dependence between objects in the scene affects the recognition accuracy of distant objects, such as the edges of the object contour, which affects its depth accuracy. We introduce non-local block [11] to our depth estimation task. Different from traditional CNN local operations obtaining the relationship between adjacent pixels, the module can calculate the interaction between any two locations through its non-local operation, directly capturing the dependencies relating to long-distance pixels.

The non-local means was first proposed to calculate the nonlocal average of all pixel points in an image [12]. Later, Wang et al. [11] proposed non-local operation, adopting a deep convolutional neural network to capture long-distance dependencies. The special operation keeps the input size of the feature map constant, calculating the response at a specific location as the characteristic weight changes for all locations of the feature maps. This process calculates all the possible location indexes of the images, i.e., global information. The self-attention module proposed by Vaswani et al. [13] is a special case of the embedded gaussian method. For a given position \( a \), that is \( \frac{1}{C(y)} = f(y_a, y_b) \). \( C(y) \) is the normalization parameter. It is then possible to calculate the softmax function for all possible location indexes \( b \):

\[
y = \text{softmax}(x^T W_\Theta y) g(y) \cdot f(y_a, y_b)
\]

is used to calculate the relationship between the feature graph of the position \( a \) and all pixels that may be associated with position \( b \). \( g(y) \) is used to calculate the position eigenvalues of the feature map \( y \). \( W_\Theta \) and \( W_y \) are the learnable weights of the input vectors \( \Theta \) and \( \phi \).

Therefore, the connection form of the vectors \( \Theta \) and \( \phi \) can be obtained: \( f(y_a, y_b) = W_\phi^T [\Theta(y_a), \phi(y_b)] \).

Among them, […] indicates the join operation, and \( W_\phi \) denotes the contact vector that can be converted into a scalar weight to enable a wider searches that capture the dependencies of long-distance pixels. Our experiments show that the non-local operation increases estimation accuracy by calculating the relationship between pixels having a specific distance and learning the similarity between pixels adaptively. Thus, it captures the long-distance spatial dependence of each pixel, reconstructing the depth information of the input feature map and obtaining more spatial context feature information.
3. Experiment
We conduct our experiments on the indoor depth estimation and evaluate the performance of our model MCEN. We implement it on the public deep learning platform PyTorch and use a NVIDIA Titan Xp GPU.

3.1. Model Setup
Our model are trained on the NYU Depth V2 dataset [14] with official split in 249 training and 215 testing indoor scenes. We adopt the SGD optimizer with an initial learning rate of 0.0001 and weight decay are set to be 0.0005.

3.2. Results and Analysis
Contribution of each component. We propose an ablation subtask to evaluate the effectiveness of each improvement in our MCEN mode, as shown in Table 1. In our method, there are three components i.e., CRE-HRNet, RM and non-local spatial module. The basic HRNet cannot obtain good result. When HRNet is replaced with CRE-HRNet, there is a slight improvement, in case of the threshold accuracy $\delta<1.25$ from 0.853 to 0.864. Next, our RM can be regarded as a depth refinement module, which improves the threshold accuracy $\delta<1.25$ from 0.864 to 0.855. Though WEM brings less performance improvement than RM, but still contributes to the final performance, improving the the threshold accuracy $\delta<1.25$ from 0.876 to 0.885. As a result, our proposed network with each component achieves state-of-the-art performance.

| Method                  | Accuracy (higher is better) | Error (lower is better) |
|-------------------------|-----------------------------|-------------------------|
|                         | $\delta<1.25^\uparrow$ | $\delta<1.25^\uparrow$ | $\delta<1.25^\uparrow$ | REL $\downarrow$ | log10 $\downarrow$ | RMS $\downarrow$ |
| Original HRNet          | 0.853                       | 0.948                   | 0.988                   | 0.151             | 0.062             | 0.549             |
| CRE-HRNet               | 0.864                       | 0.966                   | 0.991                   | 0.143             | 0.051             | 0.475             |
| CRE-HRNet + RM          | 0.876                       | 0.976                   | 0.993                   | 0.128             | 0.048             | 0.446             |
| CRE-HRNet + RM + WEM    | 0.885                       | 0.983                   | 0.997                   | 0.125             | 0.042             | 0.422             |

As shown in Table 2, we also compare the proposed method with previous state-of-the-art methods. Our method obtained evaluation accuracy indexes $\delta<1.25$, $\delta<1.25^2$, and $\delta<1.25^3$ of 0.886, 0.985, and 0.997, respectively. The results of our method are superior to other network structures, showing state-of-the-art performance.

| Method                  | Accuracy (higher is better) | Error (lower is better) |
|-------------------------|-----------------------------|-------------------------|
|                         | $\delta<1.25^\uparrow$ | $\delta<1.25^\uparrow$ | $\delta<1.25^\uparrow$ | REL $\downarrow$ | log10 $\downarrow$ | RMS $\downarrow$ |
| Eigen et al. [1]        | 0.611                       | 0.887                   | 0.971                   | 0.907             | 0.212             | **0.215**         |
| Li et al. [15]          | 0.820                       | 0.960                   | 0.989                   | 0.139             | 0.058             | 0.505             |
| Eigen et al. [4]        | 0.769                       | 0.950                   | 0.988                   | 0.158             | —                 | 0.641             |
| Laina et al. [16]       | 0.811                       | 0.953                   | 0.988                   | 0.573             | 0.055             | 0.127             |
| Hu et al. [17]          | 0.834                       | 0.968                   | 0.991                   | 0.126             | 0.054             | 0.555             |
| Our method              | 0.886                       | 0.985                   | 0.997                   | 0.119             | 0.051             | 0.407             |
Figure 2. Qualitative evaluations on the NYU Depth V2. Each line shows the results, from top to bottom, input RGB images (top), ground truth (center) and our predicted deep maps (bottom).

Compare to other state-of-the-art methods, such as the method of regressing CNN to depth information [1]. MCEN improved the accuracy by 0.275 and produced a more accurate monocular depth estimation using the dependency of the non-local block to capture distant pixels. In another case, compared to the dilated monocular depth estimation framework [15], our MCEN adopted the cascaded dilated convolution RM improving the accuracy $\delta<1.25$ by 0.066.

Qualitative results have been seen in Figure 2, our methods achieve higher accuracy and obtain smooth and delicate depth maps, effectively avoiding mesh problems that occur during feature extraction. Differing from other multi-scale CNN methods, such as Eigen et al. [4], our MCEN adopts the FP from low-dimensionality to high-dimensionality improving the accuracy $\delta<1.25$ by 0.117. This shows that MCEN retained more detailed feature information.

4. Conclusion and future work
In this paper, we propose a monocular depth estimation framework called MCEN. It uses CRE-HRNet to extract multiple scale features from monocular image. RM then obtains more contextual information. Finally, non-local block captures the depth of distant pixels. Our experimental results show that our method outperforms existing state-of-the-art methods. In future work, we plan to refine the network to achieve higher real-time depth estimation performance.

Acknowledgments
This work is supported by the Natural Science Foundation of Shanghai, China, under Grant (19ZR1435900)

Reference
[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in Advances in neural information processing systems, 2014, pp. 2366-2374.
[2] S. J. P. o. t. R. S. o. L. S. B. B. S. Ullman, "The interpretation of structure from motion," vol. 203, no. 1153, pp. 405-426, 1979.
[3] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep learning. MIT press Cambridge, 2016.
[4] D. Eigen and R. 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, 2015, pp. 2650-2658.
[5] L. Sumei, L. Guoqing, and F. J. A. O. S. Ru, "Depth Map Super-Resolution Reconstruction Based on Convolutional Neural Networks," vol. 37, no. 12, pp. 124-132, 2017.

[6] J. Xie, R. Girshick, and A. Farhadi, "Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks," in *European Conference on Computer Vision*, 2016, pp. 842-857: Springer.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5693-5703.

[8] Q. Zhou et al., "Multi-scale deep context convolutional neural networks for semantic segmentation," vol. 22, no. 2, pp. 555-570, 2019.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117-2125.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. J. I. t. o. p. a. Yuille, and m. intelligence, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," vol. 40, no. 4, pp. 834-848, 2017.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794-7803.

[12] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 2, pp. 60-65: IEEE.

[13] A. Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[14] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*, 2012, pp. 746-760: Springer.

[15] B. Li, Y. Dai, and M. J. P. R. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," vol. 83, pp. 328-339, 2018.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, 2016, pp. 239-248: IEEE.

[17] J. Hu, M. Ozay, Y. Zhang, and T. 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)*, 2019, pp. 1043-1051: IEEE.