INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.127 Volume 15, 2021

Efficient Monocular Depth Estimation with Transfer Feature Enhancement

Ming Yin,
School of Information Engineering, Xinyang Agriculture and Forestry University
No.1 North Circular Road, Pingqiao District, Xinyang, 464000.
China

Received: March 31, 2021. Revised: August 16, 2021. Accepted: August 26, 2021. Published: August 27, 2021.

Abstract- Estimating the depth of the scene from a monocular image is an essential step for image semantic understanding. Practically, some existing methods for this highly ill-posed issue are still in lack of robustness and efficiency. This paper proposes a novel end-to-end depth estimation model with skip connections from a pre-trained Xception model for dense feature extraction, and three new modules are designed to improve the upsampling process. In addition, ELU activation and convolutions with smaller kernel size are added to improve the pixel-wise regression process. The experimental results show that our model has fewer network parameters, a lower error rate than the most advanced networks and requires only half the training time. The evaluation is based on the NYU v2 dataset, and our proposed model can achieve clearer boundary details with state-of-the-art effects and robustness.

Keywords- Depth estimation, Transfer learning, Deep learning, Feature enhancement.

I. Introduction

The depth estimation of monocular images is vital for computer vision tasks, which can be applied in many fields, including detection[1], segmentation[2], intelligent control[3], and pose estimation[4]. Adequate applications in automated industry and driverless cars[5] rely on the depth estimation method to measure 3D information to achieve scene reconstruction[6]. In other words, the depth is the distance between the camera and the objects. The main job requires a solution that can make good use of the plane details, shapes and prior knowledge from two-dimensional RGB images to explore the actual three-dimensional distance.

In recent years, depth estimation methods have made some progress via deep learning due to the convolutional neural network's feature representation effect. The CNN can help to understand RGB image semantic information and translate it to an RGB-D image. Though the quantitative evaluation becomes better, the actual prediction of depth maps still has low robustness and efficiency. At some point, higher accuracy is relative because the results cannot correspond with the input images, which means the loss of origin information. Missing details or low resolution may lead to divergence and incorrect judgment of intelligent decisions for applied robots. The main problems for depth estimation are the lack of specific details and inaccuracy for even areas; Therefore, we hope to propose a novel method to optimize the depth estimation model, maintaining both high-frequency information and object boundaries. Then, the balance of the predicting effect and quantitative indicators can work well at the same time.

We analyze recent excellent CNN depth estimation models and propose a new design to make the model can not only have good quantitative results but also ensure the depth map quality. It is difficult to normalize the feature resolution from different convolutional layers for effective concatenation by skip connections. In addition, the change in resolution may lead to higher error and convergence difficulty, so our main goal is to enhance transferred features efficiency, and normalize them achieving state-of-the-art accuracy for depth estimation. To realize our goal more efficiently, we actually add some small kernel convolutions to reduce the computational amount and other nonlinear functions to improve the regression effect.

In this paper, we propose a model that exploits transfer learning to recover high-quality depth maps. At the same time, we concentrate on the monocular image depth estimation in this research, and the experimental result show our designed module availability. Comparison with the state-of-the-art demonstrates the superiority of our proposed method, which can help to address the classic problems with predicting depth maps. Our main contribution in this paper is as follows:

- We propose three effective feature normalization modules to improve the feature aggregation process at different resolutions and make the depth inference more reliable.
- The proposed efficient end-to-end model for depth estimation helps the predicting process to maintain both the efficiency and superior accuracy of the state-of-the-art model.
- We introduce the Xception network as pre-trained encoder to accelerate training. Extensive experiments on NYU v2 demonstrate the superiority of

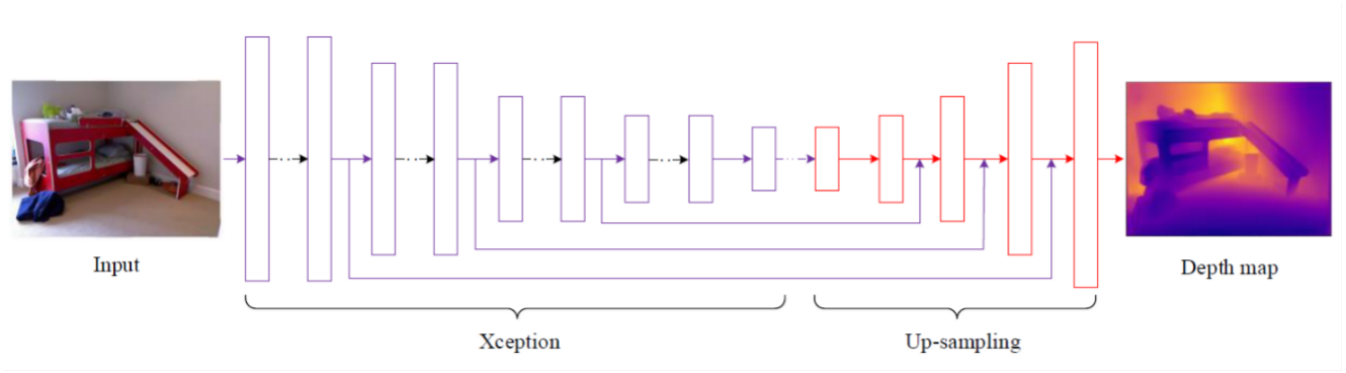

Fig. 1: Main architecture of our proposed depth estimation method.

our proposed method both qualitatively and quantitatively.

II. Related Work

In early depth estimation research, most studies were based on traditional machine learning algorithms with prior knowledge. Saxena[7] first defined a function to reflect the pixelwise relationship to a three-dimensional model and regress each pixel's depth with a Markov random field. Liu[8] proposed a discrete-continuous conditional random field (CRF) model that deeply used the pixelwise relationship and Gaussian regression to estimate the specific depth. In the same year, Ladicky[9] leveraged the theory that the image scale is inversely proportional to its depth as the core reflects, and Wang[10] proposed a nonlinear kernel function to estimate depth and obtain the kernel function parameters through sampling.

As a development of deep learning, the CNN performs well in most computer vision tasks[39, 40, 41]. Eigen[11] first introduced this network into the depth estimation field by double-scale feature fusion, which consists of coarse and refined parts. These two parts are designed to extract global and local features. Before long, he improved the network structure and fusion scale types of features[12]. Liu[13] introduced CRF loss to optimize the prediction training process, which relies on less prior knowledge. As the research became deeper, the fully convolutional neural network[14] allowed dense estimation tasks to perform better and be more widely applicable, which was combined with CRF to optimize the training results at the pixel level. For instance, Laina[15] proposed FCN with up projection and residual blocks. In detail, many deep learning pixelwise problems are based on CNN encoder-decoder models[14, 16] and have recently made some progress. Through this typical architecture, the model can well study the depth information from the input features in both supervised and unsupervised scenes. The result in our experiment also shows the superiority of the encoder-decoder architecture.

More recently, transfer learning has been applied in many neural network scenes[17]; thus, we leverage large image classification dataset pretrained models as the encoder for monocular input images. The pretrained models have great feature extraction ability[18], and some pretrained models have a great effect on maintaining spatial resolution[19]. Additionally, transfer learning is very convenient for the training process under the premise that it can not only promise the effect of the encoding but also the training efficiency. In our research, we regard transferred model selection as essential because it directly determines the quality of the captured semantic features; thus, after a series of attempts and comparisons, we finally selected the most efficient model by considering less training time and demanding dataset scope. Then, the model flexibility becomes more eco-efficient.

III. Method

A. Network Architecture

Our depth map prediction model is shown in fig. 1. As a key component for extracting different features, the frequently selected transfer learning CNN encoders include MobileNet[20] and ResNet[21]. The main criterion for selection previously was the convenience of different feature size combinations for feature normalization, which was is usually difficult.

Above all, the actual feature capture effect for the CNN model is essential, so it is inadvisable to ignore the more complex models that can also perform well. For instance, the Inception series is also a significant development route. Therefore, in our experiment, we selected the efficient Xception[22] as our encoder. Xception is an extreme Inception V3 module[23], which not only maintains accuracy with the latter but also simplifies both parameters and module architecture. Furthermore, Xception improves the traditional convolutional operation, and it is based on the hypothesis that separable convolution with both channels and spatial correlations can perform better.

Although the network can perform well in some scenes, the decoder stacked by plain convolutions limits the model effectiveness for depth estimation. Therefore, we design the modules in the decoder parts to aggregate transferred features and enhance performance. As illustrated in Fig.3, the upsampling is combined with zero-padding operations to maintain the feature resolution. This part consists of four upsampling operations with 3 types of modules. Our model leverages the ELU activa-

Table 1: Feature scale in our proposed architecture.

Input	Xception					Module a		Module b	Module c	Depth map
640×480	317×237	159×119	80×60	40×30	20×15	40×30	80×60	160×120	320×240	320×240

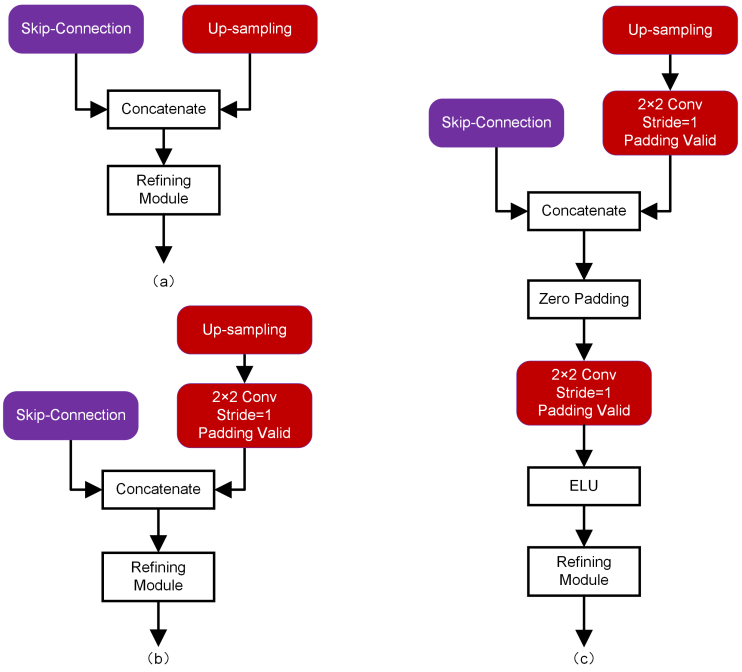

Fig. 2: The proposed feature enhancement modules.

tion function[24] as a replacement for ReLU and 2×2 and 1×1 convolutions to improve the regression process[20]. The proposed module a M_a can be formulated as follows:

$$M_a = \mathcal{F}_1(\text{Concat}(F_{\text{sc}}, \text{Upsampling}(F_d)), \quad (1)$$

where $\mathcal{F}_1$ denotes 1×1 convolution and ReLU operations, F_{sc} is the feature from the skip-connection and F_d is the features from former decoder block. The proposed module b M_b can be formulated as follows:

$$M_b = \mathcal{F}_1(\text{Concat}(F_{\text{sc}}, \mathcal{F}_2(\text{Upsampling}(F_d))), \quad (2)$$

where $\mathcal{F}_2$ denotes 2×2 convolution and ReLU operations. And the module c M_c can be formulated as follows:

$$M_c = \mathcal{F}_3(\mathcal{F}_2(\text{Concat}(F_{\text{sc}}, \mathcal{F}_2(\text{Upsampling}(F_d)))), \quad (3)$$

where $\mathcal{F}_3$ denotes ELU activation and 1×1 convolution operations.

The encoder-decoder inferring process involves changing the feature resolution, and there will be some information loss during upsampling. Otherwise, the output for the encoder and the input for the decoder are low-resolution features that enhance the pixel-level recovery difficulty. The skip-connections shown in fig.1 helps the model reserve various original details from the input RGB images and is a good strategy for image reconstruction and segmentation[2, 5, 6], such as U-Net[25] and pixel2pixel[26]. Therefore, we use differently sized features captured from Xception to directly connect to the relevant decoder layers. Through concatenation, the feature maps can be a good supplement for the encoder-decode model. Table 1 lists the feature scale in our proposed architecture, and the features of first three scales in Xception are aggregated into the decoder by the designed normalizing modules. Thus, the feature scales can be effectively unified in the decoder.

B. *Loss Functions*

The general loss function[11] reflects the depth estimation pixel-wise regression degree between the ground-truth depth image y and the predicted depth image $\hat{y}$. As an essential content, the loss function can significantly influence the actual training process, especially the model convergence speed. In our experiment, we define the depth estimation loss function $\mathcal{L}$ into two parts as follows:

$$\mathcal{L}(y, \hat{y}) = (1 - \alpha)\mathcal{L}_{pixel}(y, \hat{y}) + \alpha\mathcal{L}_{\text{MS-SSIM}}(y, \hat{y}) \quad (4)$$

The first part $L_{pixel}(y, \hat{y})$ is based on L1 regularization to calculate the divergence from the prediction to the ground-truth in the pixel value level.

$$\mathcal{L}_{pixel}(y, \hat{y}) = \frac{1}{N}\sum_{p}^{N} |y - \hat{y}| \quad (5)$$

The structural similarity (SSIM)[27] is also precise for describing the distance of two similar images, and it can perform well in unsupervised depth estimation learning[28]. Therefore, we bring in multi-scale considered SSIM in our loss function definition. Although there are

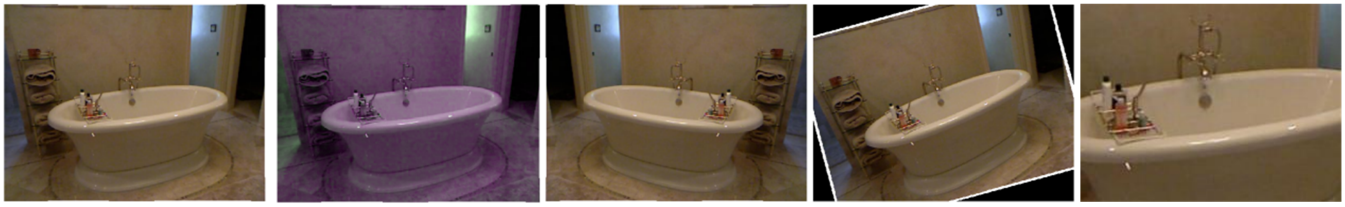

Fig. 3: Visualization of the basic data augmentation operations.

Table 2: The results comparison of our method and the state-of-the-art methods

Methods	Higher is better			Lower is better		
	δ<1.25	δ<1.25^2	δ<1.25^3	rel	log_{10}	rms
Li[31]	0.621	0.886	0.968	0.232	0.094	0.821
Liu[32]	0.650	0.906	0.976	0.213	0.087	0.759
Eigen[11]	0.611	0.887	0.971	0.215	-	0.907
Eigen and Fergus[12]	0.769	0.950	0.988	0.158	-	0.641
Chakrabari[33]	0.806	0.958	0.987	0.149	-	0.620
Laina[16]	0.811	0.953	0.988	0.127	0.055	0.573
Chen[34]	0.818	0.958	0.988	**0.123**	**0.053**	0.569
Chen[35]	0.826	0.964	0.990	0.138	0.101	0.496
Li[36]	0.832	0.965	0.989	0.134	0.095	0.540
Yan[37]	0.813	0.965	0.989	0.135	-	0.502
Xu[38]	0.811	0.954	0.987	0.121	-	0.583
ours	**0.850**	**0.973**	**0.994**	**0.123**	**0.053**	**0.461**

Table 3: The results comparison of different feature alignment operations

Methods	Higher is better			Lower is better		
	δ<1.25	δ<1.25^2	δ<1.25^3	rel	log_{10}	rms
(a)+(b)+(c)	0.850	0.973	0.994	0.123	0.053	0.461
(a)+(b)+3*(b)	0.847	0.969	0.987	0.125	0.055	0.466
ZeroPadding	0.839	0.955	0.982	0.131	0.060	0.472
ReplicationPadding	0.845	0.972	0.990	0.125	0.055	0.462

already some trials for improving the loss function[15, 16, 29], we want to improve the convergence process by employing multi-scale structural similarity (MS-SSIM)[30] to make the model reserve more high-frequency information and object details.

$$\mathcal{L}_{\text{MS-SSIM}}(y, \hat{y}) = 1 - \text{MS-SSIM}(y, \hat{y}) \tag{6}$$

C. Data Augmentation

As an important step before CNN training, the data augmentation usually makes the model more robust and avoids overfitting. In this paper, we mainly refer to how former experiments[11] pre-processed input data. In addition, swapping the color channels in the experiment can help the model learn similar images with various changes. The main pre-processing methods we use in the experiment are as follows:

- The training data are randomly rotated $r \in [-10°, 10°]$.
- Color: The training values multiply a random value c range from [0.8,1.2].
- The training pairs are horizontally flipped with 0.5 probability.
- The RGB channels of input image are randomly swapped with 0.25 probability.
- The input data are centered cropped and then resized to the former size.

IV. Experiments

A. Implementation Details

In this paper, our proposed model is trained and tested on the NYU V2 dataset[6], which includes more than 12K indoor scenes with both RGB and depth images sampled by the Microsoft Kinect camera. In our experiment, the dataset is divided into three parts for training, validation and testing. Following the official

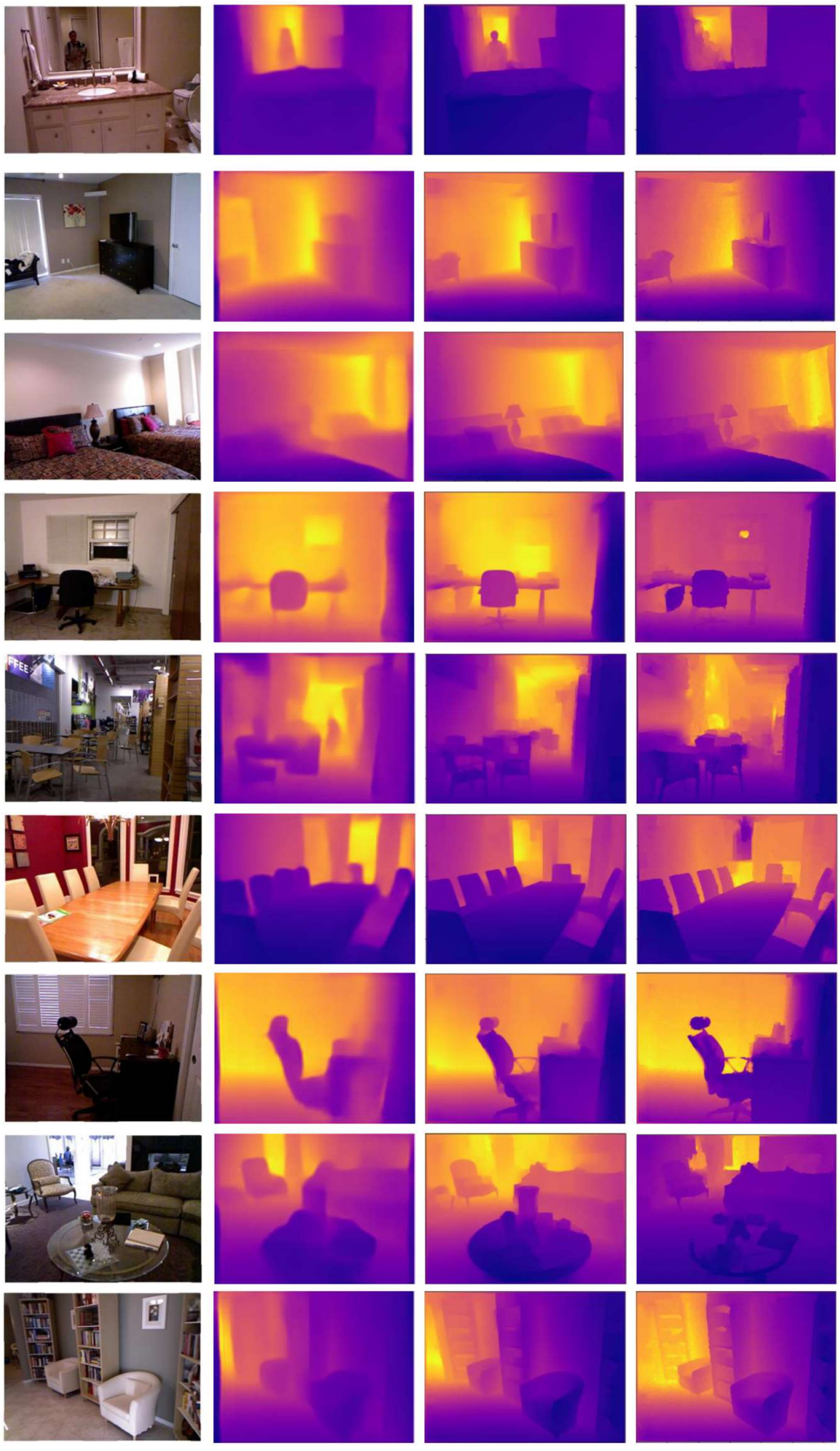

Fig. 4: Visualization of results by our method and the state-of-the art method for monocular depth estimation. (a) Input RGB image; (b) Chen *et al.* [34]; (c) Our method; (d) Groundtruth.

dataset divisions, the training part consists of 120K images, and the validation and testing parts include the same number of 659. As a pre-processing detail, the invalid area of depth groundtruth, especially the opened windows and doors that cannot be estimated, is set to the maximum value. The resolution of the input images is 640×480.

The CNN experimental environment is PyTorch, and the training hardware is based on an i7-9500 CPU, NVIDIA GTX TITAN X GPU and 128 GB of memory. The initial learning rate is 0.0001 with a 0.999 training decay. The parameters for the ADAM optimizer have a 0.0001 learning rate, 0.9 β_1 and 0.999 β_2. The ELU optimizer is 1.0 α. The preset batch-size is 8, and the epoch is 10, which requires nearly 35 hours. The final trained model is approximately 38M parameters. Additionally, the frozen weights operation for transfer learning is leveraged, so the first few layer weights that are trained for the ImageNet dataset are set to untrainable.

B. Evaluation Metrics

According to previous research details, the most commonly used quantitative evaluations are average relative error (rel), root mean squared error (rms), mean log error (log_{10}) and accuracy with three thresholds. These evaluation metrics, which follow [11], are defined as follows:

- Mean relative error (rel):

$$\sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i - g_i)}, \tag{7}$$

- Root mean squared error (rms):

$$\frac{1}{T}\sum_{i=1}^{T}\frac{||d_i - g_i||_1}{g_i}, \tag{8}$$

- Mean log error (log_{10}):

$$\frac{1}{T}\sum_{i=1}^{T}||log_{10}d_i - log_{10}g_i||_1, \tag{9}$$

- Threshold accuracy:

$$\max(\frac{d_i}{g_i}, \frac{g_i}{d_i}) = \delta{<}threshold, \tag{10}$$

where T is the number of pixels in each depth image. d_i and g_i are the prediction and ground-truth pixel-wise values, respectively.

C. Experimental Analysis

In Table 2, we show the quantitative effect by comparing the result of our method with the state-of-the-art method in terms of six evaluation metrics. The superiority of our method is obvious. Specifically, compared with competing methods[15, 34, 35, 36, 38], our method can output better threshold accuracy. Although our method is 0.02 higher than Xu[38] for rel metric, our method outperform it by a large margin in terms of other metrics. In addition, our method can reach the state-of-the-art error rate[15, 35, 36, 37]. For the highly transferred Xception model and its feature resolution modules, the predicted depth maps perform as accurate as the state-of-the-art methods, even in boundary details. Importantly, our model training only requires 35 hours, which is approximately half of the state-of-the-art training. This demonstrates our method owns higher efficiency and accuracy at the same time.

As illustrated in fig.4, visual comparison of different methods for depth estimation demonstrates that our method perform well on both local and global depth estimation qualitatively. In detail, we compare our visual results with the state-of-the-art method Chen *et al.*[34] to further validate the effectiveness of our method. For example, our method estimates sharper shape and more accurate depth map for table lamp(the third row in fig.4), and estimates depth of sofa details as accurate as the groundtruth(the last row in fig.4). Moreover, the output depth maps can accurately predict some missing parts in the NYU v2 dataset(indicated by the bounding boxes), as illustrated in fig.5. For instance, the glass in the first row, the mirror in the second row, the door in the third row, and the chair in the fourth and fifth row.

To further analyze the contribution of different modules in our method, we used different combinations of module (b) and (c) to keep the scale consistency of extracted feature maps for skip connections. With the same training hyper parameters and condition, a quantitative comparison is shown in Table 3. When using three module (b) to replace module (c), the result appeared to a little bit degrade. However, it's still has a better performance than most methods. In addition, we compare the module (b) and (c) with padding only scale alignment methods and our modules performs favorably against them. Thus, this is a great proof that the combination of our modules can help to normalize the different scale features for concatenation.

V. Conclusion

In this paper, we propose a novel encoder-decoder depth estimation model through which depth maps can be efficiently predicted through their origin RGB images. Although the proposed model uses a transfer encoder, we propose three normalization modules that help the whole model perform better by concatenating encoder features with different resolutions. The experiments are on the NYU v2 dataset, and our method performs outstandingly compared with the state-of-the-art methods and apparently demonstrates its superiority in efficiency and accuracy. Next, we plan to explore how to regularize or improve our decoder modules and improve the application efficiency to make it more easily used in embedding systems or mobile terminals.

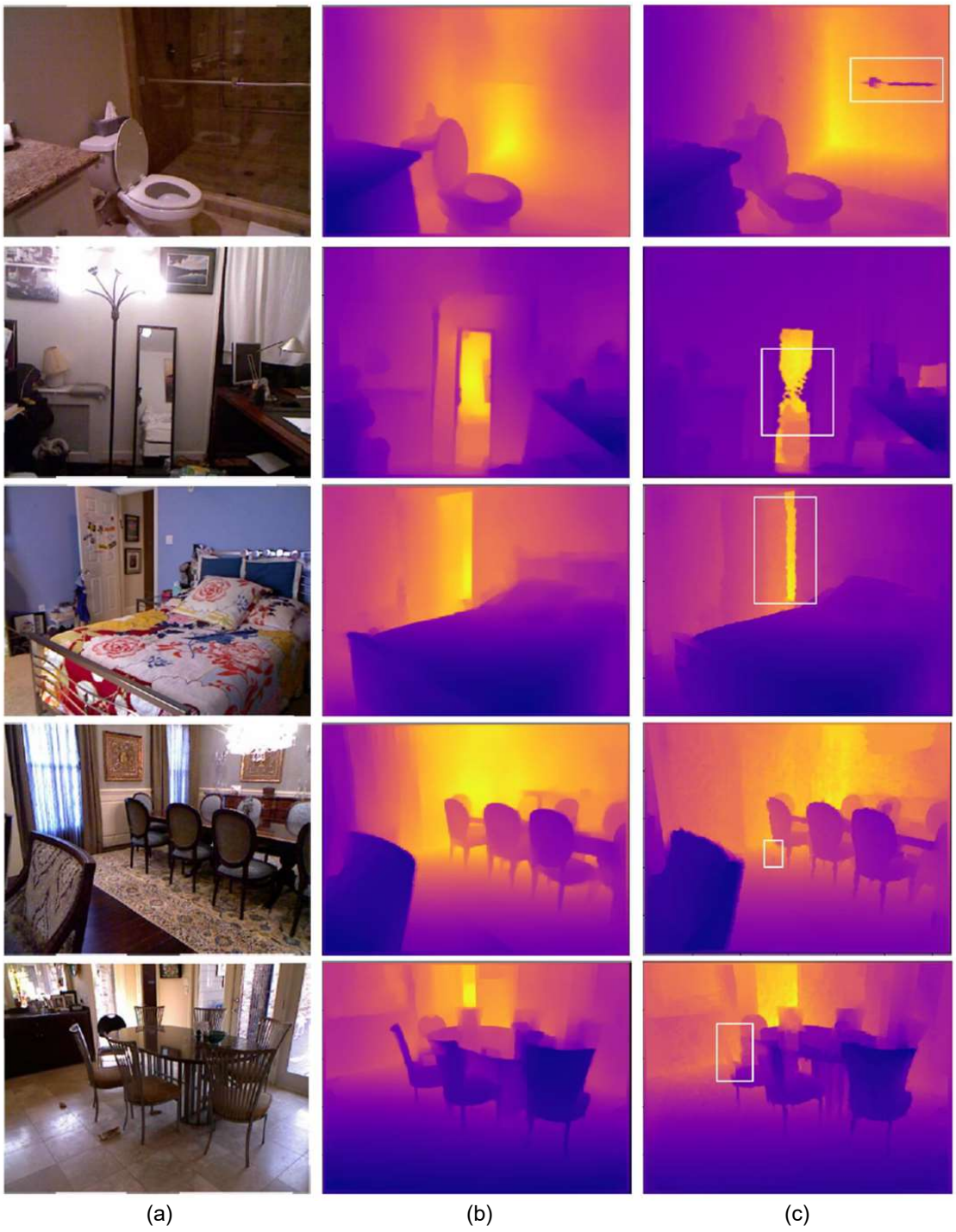

Fig. 5: The robustness of our proposed method.(a) Input RGB image; (b) Our results; (c) Error area in Groundtruth indicated by bounding boxes.

References

[1] Tang Z , Hwang J N . MOANA: An Online Learned Adaptive Appearance Model for Robust Multiple Object Tracking in 3D[J]. IEEE Access, 2019:1-1.

[2] Lian C , Ruan S , Denoeux T , et al. Joint Tumor Segmentation in PET-CT Images Using Co-Clustering and Fusion Based on Belief Functions[J]. IEEE Transactions on Image Processing, 2019, 28(2):755-766.

[3] Ragaglia M , Zanchettin A M , Rocco P . Trajectory generation algorithm for safe human-robot collaboration based on multiple depth sensor measurements[J]. Mechatronics, 2018, 55:267-281.

[4] Alp Güler R, Neverova N, Kokkinos I. Densepose: Dense human pose estimation in the wild[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7297-7306.

[5] Hong Z , Ai Q , Chen K . Line-laser-based visual measurement for pavement 3D rut depth in driving state[J]. Electronics Letters, 2018, 54(20):1172-1174.

[6] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]. European Conference on Computer Vision. Springer,

Berlin, Heidelberg, 2012: 746-760.
[7] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[C]. Advances in neural information processing systems. 2006: 1161-1168.
[8] Liu M, Salzmann M, He X. Discrete-continuous depth estimation from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 716-723.
[9] Ladicky L, Shi J, Pollefeys M. Pulling things out of perspective[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 89-96.
[10] Wang Y, Wang R, Dai Q. A parametric model for describing the correlation between single color images and depth maps[J]. IEEE Signal Processing Letters, 2013, 21(7): 800-803.
[11] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]. Advances in neural information processing systems. 2014: 2366-2374.
[12] Eigen D, Fergus R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture[C]. IEEE International Conference on Computer Vision. 2015.
[13] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5162-5170.
[14] Long J , Shelhamer E , Darrell T . Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 39(4):640-651.
[15] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]. 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016: 239-248.
[16] Ummenhofer B, Zhou H, Uhrig J, et al. Demon: Depth and motion network for learning monocular stereo[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5038-5047.
[17] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks?[C]. Advances in neural information processing systems. 2014: 3320-3328.
[18] Kornblith S, Shlens J, Le Q V. Do better imagenet models transfer better?[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2661-2671.
[19] Alhashim I, Wonka P. High Quality Monocular Depth Estimation via Transfer Learning[J]. arXiv preprint arXiv:1812.11941, 2018.
[20] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[22] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.
[23] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
[24] Djork-Arné Clevert, Unterthiner T, Hochreiter S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)[J]. Computer Science, 2015.
[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
[26] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.
[27] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.
[28] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.
[29] Fu H, Gong M, Wang C, et al. Deep ordinal regression network for monocular depth estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2002-2011.
[30] Wang Z, Li Q. Information Content Weighting for Perceptual Image Quality Assessment[J]. IEEE transactions on image processing, 2011, 20(5):1185-98.
[31] Li B, Shen C, Dai Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1119-1127.
[32] Liu F , Shen C , Lin G , et al. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10):2024-2039.
[33] Chakrabarti A, Shao J, Shakhnarovich G. Depth from a single image by harmonizing overcomplete local network predictions[C]. Advances in Neural Information Processing Systems. 2016: 2658-2666.
[34] Chen S, Tang M, Kan J. Predicting Depth from Single RGB Images with Pyramidal Three-Streamed Networks[J]. Sensors, 2019, 19(3): 667.
[35] Chen Y, Zhao H, Hu Z. Attention-based Context Aggregation Network for Monocular Depth Estimation[J]. arXiv preprint arXiv:1901.10137, 2019.

[36] Li B, Dai Y, He M. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference[J]. Pattern Recognition, 2018, 83: 328-339.

[37] Yan H, Zhang S, Zhang Y, et al. Monocular depth estimation with guidance of surface normal map[J]. Neurocomputing, 2018, 280: 86-100.

[38] Xu D, Ricci E, Ouyang W, et al. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5354-5362.

[39] G. Pradeep Kumar, B. Sridevi, Simulation of Efficient Cooperative UAVs using Modified PSO Algorithm, WSEAS Transactions on Information Science and Applications[J], Vol. 16, 2019, Art. #11, 94-99.

[40] Lucjan Setlak, Rafal Kowalik, Control Model of a Small Micro-class UAV Object Taking Into Account the Impact of Strong Wind, WSEAS Transactions on Systems and Control[J], Vol. 14, 2019, Art. #50, 411-418.

[41] Ayachi Errachdi, Mohamed Benrejeb, Adaptive Internal Model Neural Networks Control for Nonlinear System, International Journal of Electrical Engineering and Computer Science[J], Vol.2, 2020, 9-14.