Monocular Depth Estimation Using Information Exchange Network

Wen Su, Haifeng Zhang, Quan Zhou, *Member, IEEE*, Wenzhen Yang, and Zengfu Wang, *Member, IEEE*

***Abstract*— Depth estimation from single monocular image attracts increasing attention in autonomous driving and computer vision. While most existing approaches regress depth values or classify depth labels based on features extracted from limited image area, the resulting depth maps are still perceptually unsatisfying. Neither local context nor low-level semantic information is sufficient to predict depth. Learning based approaches suffer from inherent defects of supervision signals. This paper addresses monocular depth estimation with a general information exchange convolutional neural network. We maintain a high-resolution prediction throughout the network. Meanwhile, both low-resolution features capturing long-range context and fine-grained features describing local context can be refined with information exchange path stage by stage. Mutual channel attention mechanism is applied to emphasize interdependent feature maps and improve the feature representation of specific semantics. The network is trained under the supervision of improved log-cosh and gradient constraints so that the abnormal predictions have less impacts and the estimation can be consistent in high order. The results of ablation studies verify the efficiency of every proposed components. Experiments on the popular indoor and street-view datasets show competitive results compared with the recent state-of-the-art approaches.**

***Index Terms*— Depth estimation, information exchange, multi-scale context, high resolution, semantic information.**

I. Introduction

DEPTH estimation is a classic problem on computer vision and automatic driving. It is useful in various applications such as street scene understanding [1], pedestrian tracking [2], scene reconstruction [3], 3D object detection [4], [5], semantic segmentation [6] and human pose estimation [7].

Due to extremely low requirement on the sensor, depth estimation from a single monocular image is a more flexible and affordable solution for depth acquisition comparing with using stereo images [8] and video sequences [9]. However, the practical performance of monocular depth estimation retains unsatisfied in real-world applications. Typical approaches usually exploit statistically meaningful monocular cues such as perspective, texture, object sizes, locations and occlusions [10]–[12]. The inherent ambiguity of mapping an intensity or color measurement of certain scene into depth makes those approaches difficult to achieve reasonable accuracy. With the successes achieved by deep convolutional neural networks (DCNNs) in various tasks, researchers find it is a potentially practical way to estimate depth with DCNNs. These learning based approaches utilize hierarchical feature maps to regress depth values or classify depth labels. However, there are still four obstacles hindering acceptable estimation, which are motivations for our new framework: single scale context of features, neglecting of high-level semantic information, sub-sampling effects of DCNNs and inherent defects from supervision signals.

First, most of convolutional neural network (CNN) based approaches [31], [32] use features from specific convolutional layers. They represent single scale contextual information. Local context can force the prediction to align with local details. As the scale of context is increased, it encodes depth more accurately, clarifies local confusions and captures spatial configuration. Global context integrates an understanding of full scene which is crucial for cues such as object locations and spatial alignment. The observation that multi-scale context is inherently within the feature hierarchy of CNN, motivates us to propose a novel information exchange network. Multi-scale context in hierarchy interacts with that in their neighbor hierarchy in every stage. The fused contextual information enriches the hierarchical contextual diversity. The final contextual aggregation can greatly integrate local details, clarify local confusions, capture spatial configuration and understand full scene.

Second, existing CNN based approaches rarely pay attention to the high-level semantic information. In depth estimation, in addition to the low layer features of CNN (capturing edges, directions, color conjunctions and corners), the high layer features (describing semantic information) is critical to the localization, overall geometric layout and internal depth constraints. Motivating by this, when exchanging the contextual information among hierarchical features, we introduce mutual attention mechanism. We aggregate useful high-level semantic information to the high-resolution low layer features for emphasizing interdependent relations and improving the feature representation of specific semantics.

Manuscript received July 25, 2019; revised November 7, 2019, March 24, 2020, and May 22, 2020; accepted July 9, 2020. Date of publication July 28, 2020; date of current version June 2, 2021. This work was supported in part by the Science Foundation of Zhejiang Sci-Tech University under Grant 18022225-Y, in part by the Zhejiang Natural Science Fund Project of China under Grant LQ20F020001, in part by the Fundamental Research Funds of Zhejiang Sci-Tech University under Grant 2020Q014, and in part by the National Key Research and Development Program under Grant 2018YFB1004901. The Associate Editor for this article was D. F. Wolf. *(Corresponding author: Quan Zhou.)*

Wen Su and Wenzhen Yang are with the Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China (e-mail: wensu@zstu.edu.cn; ywz@zstu.edu.cn).

Haifeng Zhang and Zengfu Wang are with the Department of Automation, University of Science and Technology of China, Hefei 230016, China (e-mail: hfz@mail.ustc.edu.cn; zfwang@ustc.edu.cn).

Quan Zhou is with the National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: quan.zhou@njupt.edu.cn).

Digital Object Identifier 10.1109/TITS.2020.3008991

The aggregating weights automatically depend on the interdependencies between feature maps in same stage.

Third, sub-sampling effects of CNN always disturb the full resolution of dense prediction problem such as depth estimation and semantic segmentation. There are some techniques we can learn from semantic segmentation such as up-sampling architecture [13], decoder network [14] and atrous convolution [15]. Considering the difference between depth estimation and segmentation, transplantations seem to be very difficult to obtain satisfactory predictions [16]. However, our network can maintain prediction of high resolution. The high-resolution feature maps are refined with other hierarchical feature maps stage by stage. They receive multi-scale contextual information and high-level semantic information. Meanwhile, they broadcast their local contextual relation as well as low-level details.

Last, we also motivated by the observation that learning based approaches have inherent defects from supervision signal. The regression-based depth estimation approaches could suffer from unbalance because of randomness of depth in the foreground and huge amounts of pixels in the background. Besides, predicting error of the object point far away from camera always contributes more than that closing to camera. Depth closing to the camera will not be well estimated and accuracy of learned model will be greatly reduced. However, casting depth estimation as a multi-class classification problem has to face the issues that depth groups are ordinal and highly correlated rather than independent classes. Consequently, it usually takes a complicated loss and a great deal of computation [17]. We respect for the continuity attribute of depth and discard the complicated classification-based theory. Our network learns depth regression based on image-depth pairs. It is trained with log spatial log-cosh loss and gradient of prediction simply.

Our network is built on hierarchical feature maps of CNN. These feature maps describe hierarchical contextual information. The high-level features have rich semantic information. We aggregate the contextual information represented by the feature maps of adjacent hierarchy. It avoids the problem that the scale difference of contextual information is too large to boost the feature represent ability. The fused contextual information can enrich the diversity of contextual information. Our network aggregates the high-level semantic feature maps to the low-level feature maps adaptively with mutual attention mechanism. The semantic information useful for depth estimation is automatically selected by learning a large number of training samples. Our network gradually adds high-to-low resolution to form more stages. In every stage, we introduce three paths in the repeated fusions procedure to guarantee each representations receives information from other parallel representations over and over. Both the contextual information and semantic information are boosted stage by stage. We finally get a high-resolution feature map for predicting dense depth. The log spatial log-cosh loss and gradient of prediction are employed simply to regress depth on the high-resolution feature map. The contributions of this paper are summarized as follows:

- A general convolutional neural network maintaining high resolution for dense prediction is proposed. We use the network to solve monocular depth estimation. Our information exchange network is capable of effectively fusing multi-scale context information and mutual channel attention mechanism can utilize high-level semantic information.
- We propose log spatial log-cosh loss and gradient constraints for monocular depth estimation. They can significantly improve the accuracy of depth estimation.
- If making proper use of high-level semantic information, we prove that it can get better estimations with regression simply than complicated multi-class classification transformation. We also demonstrate that multi-scale context and high-level semantic information play an important role in depth estimation.

The remainder of the paper is organized as follows. In section II we review related recent literature. We detail our method in section III. In section IV we analyze our model with ablation experiments and assess the benchmark performance on NYU-Depth V2 and KITTI datasets. This is accompanied with a general discussion. We conclude in section V.

II. Literature Review

In this section, we roughly classified the related work into 3 categories: Geometry Depth Estimation, Deep-learning based Depth Estimation and Weakly supervised or Unsupervised Depth Estimation. Then we review deep-learning based methods from the respects of Multi-scale information fusing in CNNs, Multi-task learning in CNNs and Multi-class classification in CNNs.

A. Geometry Depth Estimation

In the past, depth estimations are realized in scene understanding. Most of them are often based on geometric or attributed constraints. Kim *et al.* [18] construct conditional random fields (CRFs) over a 3D volume of interest which captures the semantic and geometric relationships among different voxels of the scene. Gupta *et al.* [19] apply global geometric constraints between 3D volumes, the laws of statics on representations where objects have volume and mass, and relationships describe 3D structure and mechanical configurations. Liu *et al.* [20] construct a hierarchical parse graph by recursively applying five grammar rules while preserving the attributes constraints. These approaches are limited to modeling particular scene structures. They are not applicable for general-scene depth estimations. More recently, non-parametric methods have been explored. Karsch *et al.* [21] perform feature-based matching to search image candidates. Retrieved candidates are then warped and combined to produce the final depth map. These approaches rely on the assumption that similarities between regions in the color images imply also similar depth cues. Some researchers have found the probabilistic graph models provide a new solution to depth estimation. For example, Saxena *et al.* [10], [12], [22] employ Markov Random Fields (MRFs) to learn the parameters in supervised way. Liu *et al.* [23] treat depth estimation as inferring maximal posterior probability of CRFs. There are some approaches that combine depth estimation with other tasks. To name a few, Wang *et al.* [24] and Ladicky *et al.* [25] solve depth estimation and semantic segmentation at the same time and find they can promote each other. Liu *et al.* [26] find that semantic segmentation can improve the performance of depth estimation. Even though these efforts have contributed to monocular depth estimation, our work is based on deep learning which have greatly advanced the accuracy of single image depth estimation.

B. Deep-Learning Based Depth Estimation

Due to the successes of depth learning, researchers try to solve the monocular depth estimation with the theory of DCNNs. Laina *et al.* [16] employ fully convolutional architecture with residual learning and BerHu loss to model the mapping between monocular images and depth. The up-sampling of feature maps are learnt through residual un-pooling decoder network. Li *et al.* [27] use a two-streamed CNN to predict depth and depth gradients and then fuse them together into detailed depth map. Liu *et al.* [28] propose deep convolutional neural field model. It jointly explores the capacity of DCNN and continuous CRFs. Xu *et al.* [29], [30], Li *et al.* [31] and Cao *et al.* [32] also employ CRFs to refine the prediction generated by CNN. Roy and Todorovic [33] combine random forests and CNN to construct neural regression frost. Scanning windows extracted from images are passed down the trees and filtered with a CNN tree node to predict their depth. Zhang *et al.* [46] propose a progressive hard-mining network framework to preserve the cross-border details of depth maps. Unlike previous methods [29]–[32], [46], we focus on the role of multi-scale contextual information and semantic information in depth estimation. We therefore review the most related literature of deep-learning based depth estimation from the following aspects.

Multi-scale information fusing in CNNs Some approaches pay attention to the importance of multi-scale information. Eigen *et al.* [34] Eigen and Fergus [35] progressively generate features and refine prediction to high resolution using a sequence of three scales with a single multi-scale convolutional architecture. Xu *et al.* [29] fuse representations derived from multiple output layers of CNN based on continuous CRFs. Later, Xu *et al.* [30] improve their work with a structured attention model. It regulates the amount of information transferred between corresponding features at different scales and it is seamlessly integrated into the CRF. Fu *et al.* [17] apply dilated convolution with multiple dilation rates to extract multi-scale information and develop a full-image encoder capturing image-level information. Kim *et al.* [47] propose a deep variational model that integrates heterogeneous predictions from global and local networks. In contrast to previous approaches [29], [30], [34], [35], [47], our information exchange network allows contextual information at adjacent scale to exchange and enrich the diversity of hierarchical contextual information. Our work not only combines multi-scale context information, but also pays more attention to the utilization of semantic information. In addition, instead of using complex probability graph model to refine depth map, we only use simple and effective regression to calculate depth map.

Multi-task learning in CNNs There are approaches which fall back on multi-task learning to solve the depth estimation problem. Eigen and Fergus [35] address depth estimation, surface normal and semantic segmentation in a single multi-scale CNN. They argue usage of a single architecture helps simplify the implementation of systems that require multiple modalities. Wang *et al.* [24] observe depth estimation and semantic segmentation are strongly correlated and mutually benefit. They utilize a pre-trained CNN to predict global layout and decompose image into local segments as well as semantic prediction. Then they formulate the inference problem in hierarchical CRFs. Li *et al.* [31] use DCNN model to learn the mapping from image patch to depth and surface normal values at super-pixel level. Then the depth and surface normal values are refined to pixel level through optimize potentials on them. Jiao *et al.* [49] propose a synergy network to automatically learn the information sharing strategies between semantic segmentation and monocular depth estimation. They use the long tail property to propose an attention driven loss for the network supervision. In our approach, we focus on monocular depth estimation and we note that the relationship between semantic segmentation and depth estimation mainly lies in semantic information. Our proposed mutual attention module can automatically aggregate semantic information.

Multi-class classification in CNNs Recently, popular approaches recast depth estimation as multi-class classification. Cao *et al.* [32] formulate depth estimation as a spatially dense prediction task. They propose information gain loss to train deep fully convolutional residual network. It can obtain the confidence of depth prediction. Fu *et al.* [17] discrete depth values with spacing-increasing discretization strategy. Then they recast depth estimation as multi-class classification problem with ordered information between labels. The ordinal loss is used to learn network parameters. These methods usually take a complicated loss and a great deal of computation while we obey continuous attribute of depth. We use simply and effective log spatial log-cosh loss and gradient of prediction to regress the depth.

C. Weakly Supervised or Unsupervised Depth Estimation

Expecting deep-learning based depth estimation, there are approaches which try to use weakly supervised or unsupervised way. Kuznietsov *et al.* [36] estimate depth in semi-supervised way. They use sparse ground-truth depth for supervised learning and enforce their network to produce photo-consistent dense depth maps in stereo setup using direct image alignment loss. Garg *et al.* [37] generate an inverse warp of target image under the training of convolutional encoder with a pair of images and known camera motion. Then they reconstruct the source image using the warp. Godard *et al.* [38] exploit epipolar geometry constraints to generate disparity images. They enforce consistency between the disparities produced relative to both the left and right images to train their network. Chen *et al.* [48] propose unsupervised SceneNet to model the geometric structure of

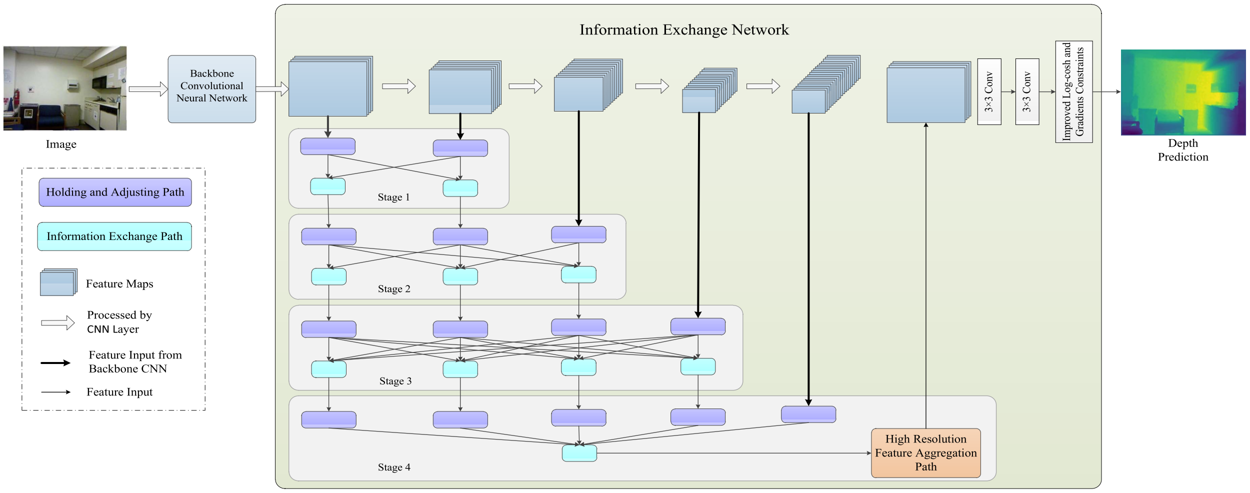

Information Exchange Network
Image
Backbone Convolutional Neural Network
Holding and Adjusting Path
Information Exchange Path
Feature Maps
Processed by CNN Layer
Feature Input from Backbone CNN
Feature Input
Stage 1
Stage 2
Stage 3
Stage 4
High Resolution Feature Aggregation Path
3×3 Conv
3×3 Conv
Improved Log-cosh and Gradients Constraints
Depth Prediction

Fig. 1. Overall network architecture. Our network is constructed on the hierarchical feature maps of backbone CNN though multi-scale contextual feature path. Multi-scale contextual feature path actually is the output feature maps of five different blocks of backbone before activation. The whole network is composed of four stages vertically. The first three stages include two kinds of processing path: holding and adjusting path, information exchange path. The last stage has an extra high-resolution feature aggregation path. The stage and other three paths are detailed in Figure.2.

objects with the aid of semantic understanding from segmentation. Their model performs region-aware depth estimation by enforcing semantics consistency between stereo pairs. Ji *et al.* [39] propose a semi-supervised adversarial learning framework to utilize a small number of image-depth pairs in conjunction with a large number of easily-available monocular images. They use one generator to regress the depth and two discriminators to evaluate the predicted depth. Ramirez *et al.* [50] propose a semi-supervised deep learning approach aimed at joint semantic segmentation and depth estimation. They leverage on semantic labels with unsupervised signals gained by geometry through an image warping loss. These methods estimate monocular depth rely on stereo setup instead of single image.

III. Method

In this section, we first detail the architecture of proposed information exchange network. Then, we discuss our innovations. The overall architecture is shown in Figure.1. Our network is constructed on the hierarchical feature maps of CNN. It consists of four stages vertically. The first three stages are composed of holding and adjusting path as well as information exchange path. The last stage has an extra high-resolution feature aggregation path. The structures of stage and paths are detailed in Figure.2. Same component is indicated with same color. Such architecture effectively aggregates multi-scale contextual information. The introduced mutual attention module enables the network to independently superimpose beneficial semantic information. The loss function can effectively avoid the influence of improper contributes of estimation and constrain the gradients of depth during regression.

A. Information Exchange With High Resolution

The superior of CNN is the feature hierarchy. The features of bottom layer are drawn from the features of upper layer through operations such as convolution, non-linear activation or pooling. The features obtained through the convolution of the low layers only reflect the local statistical characteristics within the neighborhood covered by convolutional kernel. As the operations are repeated, contextual scale (receptive field) is broadened. Larger receptive fields can describe a wider range of contextual information. Multi-scale contextual information can be complement to local features. For example, relying only on local information to determine the depth of a region on the table may be inaccurate due to the influence of color and texture. But if the network could refer to the horizontal smoothing geometric properties of the entire table, the depth estimation will not be affected by local appearance. Consequently, multi-scale contextual information plays an important role in reducing local confusions and providing spatial layout. The receptive field of the extracted feature map before aiming to specific task is generally thought to have covered the entire image area. It reflects the global contextual information. Global contextual information can provide overall cues such as geometric layout, position, outline and so on. If the features only focus on the local region of object and do not describe the relative position and geometric attributes between the objects, then depth estimation obviously will be inaccurate. Meanwhile, it is worth noting that the resolution of the prediction obtained by the network is always decreased with the increasing of the scale of context. The output resolution of the last convolution block of ResNet [40] is only $1/32$ of the input image resolution.

In view of the above reasons, the proposed network in this paper chooses ResNet-101 as backbone. The backbone generates intermediate feature maps of different scale recurring to its residual blocks which we call as multi-scale contextual features path. The proposed network is constructed based on the features of five different scales. The overall architecture is shown in Figure.1. In the figure, we fold up the details of the path. In the entire network, we take four stages as affordable point to balance enriched multi-scale information

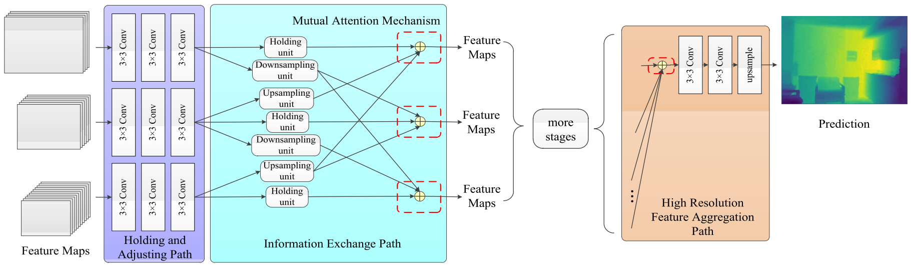

Feature Maps
3×3 Conv
Holding and Adjusting Path
Mutual Attention Mechanism
Holding unit
Downsampling unit
Upsampling unit
Information Exchange Path
Feature Maps
more stages
3×3 Conv
upsample
High Resolution Feature Aggregation Path
Prediction

Fig. 2. The stage is composed of four paths. Multi-scale contextual feature path generates five feature groups. Holding and adjusting path uses convolution to hold resolution and adjust feature channels. Information exchange path has three kinds of unit. High-resolution feature aggregation path generates final feature maps for regression. The details of information exchange path please refer to Figure.3.

and computing resources. One typical stage is illustrated in Figure.2. In every stage, each feature group is superimposed by another four feature groups and maintain its own resolution through information exchange path. Taking the first feature group as an example, the feature map operated by holding and adjusting path and the other four feature groups processed by the up-sampling unit are accumulated as an input feature group to the next stage. Each output feature group of such stage aggregates multi-scale contextual information.

Multi-scale contextual feature path takes the output feature maps of five different blocks of backbone before activation as the input feature groups. The resolutions of feature maps in these five groups are gradually decreased with strides $\{4, 8, 16, 32, 32\}$. The smaller their resolution is, the larger contextual scale is. Therefore, the choice of such input feature groups is natural. We try to aggregate feature maps effectively while maintaining a rather large resolution.

Holding and adjusting path is a preparation for the aggregation of feature maps. First, the feature channels of input feature groups are incompatible, such as $\{64, 256, 512, 1024, 2048\}$ respectively. Their spatial resolutions are also inconsistent. In each group, we aim to maintain the original resolution. Second, the feature maps with more feature channels are actually calculated on them with less channels, and the abstract of information is of varying degrees. It is unenforceable to aggregate different feature groups directly. To make the multi-scale contextual information diffusion more uniform, we use three convolution layers as holding and adjusting unit. Each feature group is processed by it before and after aggregation.

Information exchange path is used to realize the aggregation of multi-scale information. It has three kinds of units. The details of units used in Figure.2 are illustrated in Figure.3. The up-sampling unit includes convolutional layer and bilinear up-sampling layer. It can transform any input low-resolution feature map into the target resolution feature map without feature channels limit. The holding unit is composed of a convolutional layer. It can transform any input feature channels into target feature channels. The down-sampling unit is composed of a convolutional layer with stride 2. It can transform any high-resolution feature maps of any channels into target low-resolution feature maps of specific channels.

High-resolution feature aggregation path is used to obtain multi-scale contextual feature maps that can predict the depth directly. We make use of the highest resolution of feature groups (1/4), which has been maintained in the information exchange network. In the fourth stage, we add the other four feature groups to the high-resolution feature group after the up-sampling unit. After two convolutional layers, the resolution of the original image is obtained by bilinear up-sampling.

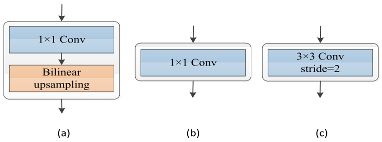


Fig. 3. (a) Up-sampling unit. (b) Holding unit. (c) Down-sample unit.

When an image is fed into the backbone network, we obtain hierarchical feature maps from multi-scale contextual feature path. These feature maps contain hierarchical contextual information and high-level feature maps have rich semantic information. In every stage, the input feature maps of adjacent hierarchy are made compatible with holding and adjusting path. Then in information exchange path, we uniform their spatial resolution and sum up multiple scale contextual information. Meanwhile we aggregate the high-level semantic feature maps to the low-level feature maps adaptively with mutual attention mechanism. Feature maps, which combine the context and semantics of adjacent scales, act as input feature map for the next stage. Both the contextual information and semantic information are boosted stage by stage. In intermediate stages, fused contextual information can enrich the diversity of contextual information. The scale difference of contextual information will not be too large to boost the feature represent ability. In our network, feature map of a specific resolution shares its own context and semantics with other feature maps in the same stage and is affected by other context and semantics. That is why we call our architecture as information exchange network.

B. Learn Meaningful Semantics With Mutual Attention

In the above-described information exchange network, the contextual scale contained in feature groups of different resolution is distinguishable. At the same time, the semantic information described is also different. According to the abstract degree of the combination between semantic elements in image, the semantic information can be divided into six levels, which is called hierarchical semantic model [41].

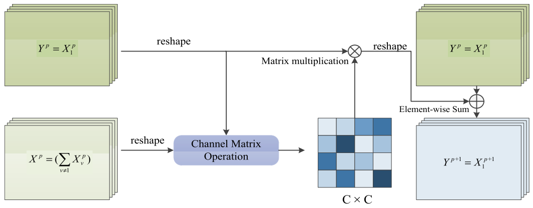


Fig. 4. Mutual channel attention module.

The low-level semantic information is also referred to feature semantics. It mainly describes the structured perception of target components in a set of connected pixels. In CNN, it generally corresponds to shallow layer features, mainly describing edges, directions, corners and so on. What is further abstracted based on low-level semantic information is high-level semantic information. The high-level semantic information can be divided into object semantics, spatial relationship semantics, scene semantics, behavior semantics, and emotion semantics, depending on the described focus of the high-level semantic information. Object semantics describe a set of structured connecting object components. For example, in the higher layers of CNN, the structure comprised of edge and texture pattern are described. Scene semantics and spatial relationship semantics are formed on the basis of object semantics. Both of which describe a structured object combination or set having a certain spatial relationship. For example, features of the higher layers of DCNNs describe the spatial locations, outlines and other information of the objects. On the basis of scene semantics and spatial relationship semantics, behavior semantics and emotion semantics are further abstracted. Features extracted by simple CNN mainly focus on feature semantics, object semantics, spatial relationship semantics and scene semantics. So far, the effect of different levels of semantics on depth estimation is not made a thorough inquiry. Empirically, not all levels of semantics is useful for depth estimation while some of them do play important roles in varying degrees. In the proposed network, with the aggregation of multi-scale contextual information, in fact, semantics are merged constantly by summation along the feature channels. Each channel of feature maps is associated with semantic information expressed by this feature map. To reasonably use semantic information, motivated by [17], [42], we introduce mutual channel attention mechanism to our network. By exploiting the interdependencies between feature groups along feature channels in same stage, we could emphasize interdependent feature maps and improve the feature representation of specific useful semantics.

The structure of mutual channel attention module is illustrated in Figure.4. In every feature group u of stage p, we use channel matrix operation to calculate the mutual channel attention map from the original feature group $Y^p = X_u^p \in R^{B\times C\times H\times W}$ and additional feature groups $X^p = (\sum_{v\neq u} X_v^p) \in R^{B\times C\times H\times W}$. Where B is the batch size, C is the number of feature channels, H and W are the height and width of feature maps respectively. Specifically, we reshape Y^p and X^p to $R^{B\times C\times (H\times W)}$ and then perform a matrix multiplication between Y^p and the transpose of X^p. The channel attention map $\alpha \in R^{(B\times C)\times(B\times C)}$ is obtained by applying softmax to the result of matrix multiplication.

$$\alpha_{ji} = \frac{\exp(X_i^p \cdot Y_j^p)}{\sum_{i=1}^{c} \exp(X_i^p \cdot Y_j^p)} \tag{1}$$

where α_{ji} is the item of matrix α. It measures the impact of ith channel of additional feature groups on the jth channel of original feature groups. In other words, the matrix is a interdependency model. It measures the influence which semantic information of additional feature groups exerts on semantic information of original feature groups. In addition, we perform matrix multiplication between the transpose of α and Y^p, and reshape their result to $R^{B\times C\times H\times W}$. This operation will give us the meaningful semantic information in additional feature groups. Then we use a scale parameter λ multiplying the result to control the ratio of contributions which meaningful semantic information makes to original feature groups. Where λ gradually learns a weight from 0. We perform an element-wise sum operation with Y^p to obtain the final output Y^{p+1}.

$$Y_j^{p+1} = \lambda \sum_{i=1}^{c} (\alpha_{ji} Y_i^p) + Y_j^p \tag{2}$$

The equation shows that the final feature of each channel is a weighted sum of the features of all channels and original feature groups, which models semantic dependencies between original feature groups and additional feature groups. Our mutual attention calculates the interdependencies between feature channels of original feature map and additional feature maps in same stage. It can be regarded as a kind of automatically regulation of semantics. The meaningful semantic information of additional feature groups is boosted with high weights while semantic redundancies are weakened with small weights adaptively.

C. Loss Function

In depth estimation, mean absolute error (MAE) and mean square error (MSE) are usually used as loss functions to supervise the training of DCNNs because of the continuity of depth values. However, the distribution of depth has a certain range. Even though the larger depth value is estimated more accurately, the error of it will still be larger than that of the smaller depth value estimated less accurately. Consequently, the contributions of more accurate estimations to loss function would be greater than the inaccurate estimations. It will be harmful for the training. Besides, the amount of depth of scene is randomness. MSE is sensitive to larger depth while MAE is sensitive to small depth. Using either MAE or MSE will surfer from slow convergence and unsatisfactory results. To solve these problems, we map the depth value to the log space. The error of small depth value has same order of magnitude as that of large depth value in log space. We make use of the inverse ratio property of the gradient in the log space so that the larger depth is also comparable to the smaller depth. We use the log-cosh loss function motivated by [16]. We set $Dif_{\log} = \left|\log_{10} D_g\right| - \left|\log_{10} D_p\right|$, where D_p is the depth estimation, D_g is the ground-truth. The training dataset have N images and i is the index of image. The loss function is formulated as follows.

$$B = \sum_{i=1}^{N} \log(\cosh(Dif_{\log})) \tag{3}$$

When the error in log space is small, $log(cosh(x))$ is approximately equal to $x^2/2$. The loss function optimizes the network with L2 norm. When the error in log space is big, $log(cosh(x))$ is approximately equal to $|x|-log2$. It is a kind of L1 norm. The network will be optimized by L1 norm. Thanks to this piecewise property, abnormal predictions do not have that much impact on predictions. Log-cosh loss shows a good balance when the residual is varying. It puts high weight towards pixels with a small residual because of the L2 term. At the same time, L1 accounts for a decreased impact of high residuals gradients than L2 would. In addition, in order to make the depth gradient consistent with the ground-truth in high order, we also constrain the gradients of estimations. The gradients constraints and global loss of network are formulated as follows. Where $D_g{}'$ and $D_p{}'$ are the gradients of ground-truth and depth prediction respectively.

$$G = \sum_{i=1}^{N} \left|D_g{}' - D_p{}'\right| \tag{4}$$

$$L = B + \beta G \tag{5}$$

We use a parameter β to balance the weights of two parts. Where β usually is set empirically according to specific dataset. We empirically set β in the range $[5, 10]$. In our experiments we also find that it is not an important parameter for boosting the performance excepting an influence on convergence rate. The final loss can force the prediction to better process random depth values and align with ground truth boundaries.

IV. Experiments

In order to verify the efficiency of our proposed network, we carried out various ablation studies and comparative experiments with other SOTA methods. In this section, we first briefly introduce the datasets, quantitative evaluation metrics and the setting of hyper parameters in training phrase. Then we prove the importance of semantics to depth estimation. We explore the effect of the number of feature channels in information exchange network. We also prove the effectiveness of adding mutual attention mechanism to the information exchange network and using our log space gradient loss function. Finally, we compare the proposed approach with other SOTA methods to demonstrate the overall effectiveness.

A. Datasets, Metrics and Experimental Setting

We have evaluated our proposed monocular depth estimation method on both indoor and outdoor datasets: NYU-Depth V2 and KITTI. The NYU-Depth V2 [43] dataset is comprised of video sequences from various indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It has 1449 densely labeled pairs of aligned RGB and depth images. Its raw dataset has 464 scenes taken from 3 cities and is officially split into 249 training and 215 test scenes. When we evaluate our entire architecture, we sample equally-spaced frames out of each training sequence, resulting in approximately 90k unique images. We train our model using this raw training data and test on the standard 654 test images which is same as [24], [28], [33], [34], [16], [32], [17]. KITTI dataset [44] is composed of several outdoor scenes captured while driving with car mounted cameras and depth sensor. We train our model on the same training set in [38] which contains 33131 images and test on the same 697 images in [28], [34], [17], [32], [36], [38]. For the implementation of our network, we use pytorch framework and train on two NVIDIA GeForce GTX 1080Ti with 11GB of GPU memory. We use Stochastic Gradient Descent (SGD) to train our network. Our network is initialized from the model ResNet-101 pretrained on the ILSVRC for image classification. All newly added layers are initialized as normal distribution with zero mean and 0.01 variance. The learning rate of network is set to 0.001. The weight decay and the batch size are set to 0.0005 and 6 respectively. The β is set to 10 for NYU-Depth V2 and 8 for KITTI empirically. For quantitative evaluations, we use the same measures commonly used in prior works [31], [33], [32], such as root mean squared error (RMSE), average relative error (REL), average log10 error (LOG10), root mean squared log error (RMSELOG) and accuracy with threshold (δ, δ^2, δ^3). The definitions of quantitative evaluations are as follows. Where T is total number of pixels in all evaluated images and t is the index of pixels. d_g and d_p are the pixels in D_g and D_p respectively.

$$RMSE = \sqrt{\frac{1}{T}\sum\nolimits_t \left(d_g - d_p\right)^2} \tag{6}$$

$$REL = \frac{1}{T}\sum\nolimits_t \frac{\left|d_g - d_p\right|}{d_g} \tag{7}$$

$$LOG10 = \frac{1}{T}\sum\nolimits_t \left|\log_{10}d_g - \log_{10}d_p\right| \tag{8}$$

$$RMSELOG = \sqrt{\frac{1}{T}\sum\nolimits_t \left(\log_{10}d_g - \log_{10}d_p\right)^2} \tag{9}$$

$$\text{Percentage of } d_p \text{ s.t. } \max\left(\frac{d_g}{d_p}, \frac{d_p}{d_g}\right) < \delta \tag{10}$$

B. Ablation Studies

For ablation studies, we use the standard training set containing 795 images of the NYU-Depth V2 labeled dataset and evaluate on the standard 654 test images.

1) Is High-Level Semantic Information Useful in Depth Estimation: In order to explore the semantics in CNN, we visualize hierarchical feature maps. As illustrated in Figure.5. The work of [45] provides a visualization technique which reconstructs intermediate features in image space to characterize CNN. We apply the inversion technique to analyze the semantic information of backbone network. Projections from different layers show the hierarchical characteristics of the feature maps in the network. Lower layers (Maxpooling and Block1) respond to edge/color conjunctions and corners.

TABLE I

DEPTH ESTIMATION FROM DIFFERENT FEATURE MAPS OF CERTAIN LAYER WITH L2 LOSS. WE ESTABLISH BASELINES BY TAKING SPECIFIC FEATURE MAP AS THE INPUT OF THE CONVOLUTION LAYER AND THE UP-SAMPLING LAYER, AND OBTAIN THE DEPTH REGRESSION WITH L2 LOSS

Methods	Error				Accuracy		
	REL	LOG10	RMSE	RMSELOG	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
Maxpooling(baseline1)	0.4836	0.1635	1.3037	0.2093	0.3874	0.6636	0.8332
Block1(baseline2)	0.5176	0.1674	1.3091	0.2134	0.3767	0.6505	0.8235
Block2(baseline3)	0.3997	0.1436	1.1701	0.1841	0.4296	0.7241	0.8834
Block3(baseline4)	0.3762	0.1339	1.0840	0.1727	0.4582	0.7600	0.9043
Block4(baseline5)	**0.3128**	**0.1258**	**1.0359**	**0.1597**	**0.4703**	**0.7835**	**0.9288**

TABLE II

DEPTH ESTIMATIONS WITH DIFFERENT NUMBER OF FEATURE CHANNELS WITH GRADIENT CONSTRAINTS AND L2 LOSS

Channel configuration	Error				Accuracy		
	REL	LOG10	RMSE	RMSELOG	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
$\{64, 256, 512, 1024, 2048\}$	0.1908	0.1750	0.6649	0.2396	0.7274	0.9303	0.9809
$\{64, 256, 512, 512, 512\}$	0.2026	0.1844	0.6782	**0.2134**	0.7006	0.9216	0.9806
$\{512, 512, 512, 512, 512\}$	0.1850	0.1698	**0.6417**	0.2337	0.7391	0.9339	0.9826
$\{256, 256, 256, 256, 256\}$	**0.1826**	**0.1694**	0.6436	0.2321	**0.7404**	**0.9362**	**0.9835**
$\{128, 128, 128, 128, 128\}$	0.1868	0.1735	0.6617	0.2365	0.7295	0.9322	0.9834
Multi-class FRCN[32]	0.1920	0.0770	0.6880	–	0.7220	0.9260	0.9800

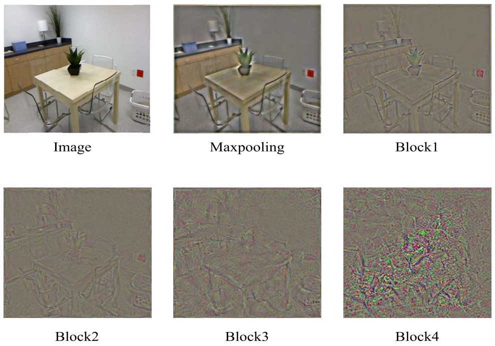


Fig. 5. Visualization of features of backbone network inverted to image space. The upper left corner is the original image. Maxpooling is generated by the first max-pooling layer. Block i is generated by the ith block of backbone before non-linear activation.

Intermediate layers (Block2 and Block3) have more complex invariance, capturing similar textures and structures. High layers (Block4) show significant variation, and is more class-specific. They show entire objects with significant pose variation. To compare the influence of different levels of semantics to depth estimation, we use the features of a specific layer alone for depth estimation. We take specific feature map as the input of the convolution layer and the up-sampling layer, and obtain the depth regression with L2 loss. The results are shown in Table.I, which are also our baselines.

From the results of Table.I, we can draw conclusion that larger scale of context can contribute to depth estimation. Features from high layers always get finer estimation. Because the feature maps from block4 and feature maps from Block3 have the similar receptive field, the better performance of Block4 shows that high-level semantic information is indeed more beneficial to depth estimation. However, the estimation from Block4 has long way from satisfactory results. It shows that single scale context and only high-level semantic information are obviously not enough for accurate depth estimation. Besides, the improvements from every layer are different. It shows that some of the high-level semantic information is useful but it may introduce adverse effects in the process of gradual abstraction.

2) Feature Channels in Information Exchange: In the experiment, we find the number of feature channels in the information exchange network is an important factor affecting the regression effect. Therefore, we try to select different feature channels. The gradient constraints and L2 loss are used as the loss function to train the network. The results with different number of feature channel are shown in Table.II.

We can draw two conclusions from the results. The third row gets the best estimation. Each channel of feature maps is associated with semantic information. We think the adjustment of the interactive channels of the information exchange network is a kind of adjusting the high-level semantic information in the feature aggregation. Compared the results of first three rows, we can find that either more high-level semantic information or more low-level details can contribute to estimation. The best performance is achieved when the feature channels are equal. It means depth estimation depends on semantic information and low details at the same time. Besides, the third row and the fourth row are better than the last row which is listed in [32] which use the same backbone network. It shows that choosing proper feature channels the depth estimation of simple regression can perform better than that of multi-class classification. We think this is an interesting heuristic for depth estimation.

Actually, we also conduct the experiments of passing the features of a specific layer to the decoder similar to FCN [13]. We pass the features at different resolutions of backbone to the decoder of FCN-32s, FCN-16s and FCN-8s. The results are shown in Table.III. The gradient constraints and L2 loss are

TABLE III

DEPTH ESTIMATIONS OF PASSING THE FEATURES OF A SPECIFIC LAYER TO THE DECODER SIMILAR TO FCN

Methods	Error				Accuracy		
	REL	LOG10	RMSE	RMSELOG	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
Backbone+De-FCN32s	0.4564	0.1453	1.2797	0.1803	0.3479	0.6329	0.8012
Backbone+De-FCN16s	0.3451	**0.1028**	1.0539	0.1419	0.4327	0.7330	0.8739
Backbone+De-FCN8s	0.2928	0.1043	1.0138	**0.1289**	0.4501	0.7581	0.9072
Ours(Feature channel-256)	**0.1826**	0.1694	**0.6436**	0.2321	**0.7404**	**0.9362**	**0.9835**

TABLE IV

DEPTH ESTIMATION WITH ATTENTION MECHANISM. WOCA – WITHOUT CHANNEL ATTENTION. SCA – WITH SELF-CHANNEL ATTENTION. MCA – WITH MUTUAL CHANNEL ATTENTION

	Error				Accuracy		
	REL	LOG10	RMSE	RMSELOG	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
WOCA	0.1908	0.1750	0.6649	0.2397	0.7274	0.9303	0.9809
SCA	**0.1870**	0.1747	0.6618	0.2377	0.7271	0.9316	0.9827
MCA	0.1872	**0.1734**	**0.6564**	**0.2366**	**0.7293**	**0.9332**	**0.9830**

TABLE V

DEPTH ESTIMATION RESULTS OF DIFFERENT LOSS WITH FOURTH CHANNEL CONFIGURATION

	Error				Accuracy		
	REL	LOG10	RMSE	RMSELOG	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$
L2	0.2081	0.1970	0.7243	0.2614	0.6644	0.9118	0.9787
Log-cosh	0.2055	0.1927	0.7189	0.2578	0.6775	0.9170	0.9777
Berhu[16]	0.1998	0.1834	0.6839	0.2496	0.7052	0.9214	0.9799
Log spacial Log-cosh	0.1879	0.1743	0.6639	0.2375	0.7252	0.9331	0.9830
Log spacial Log-cosh + Grad	**0.1814**	**0.1739**	**0.6358**	**0.2315**	**0.7453**	**0.9369**	**0.9844**

used as the loss function to train the network. In contrast to FCN that sequentially merges predictions from middle layers to final segmentation output, our information exchange network explores different stage features in a hierarchical manner, resulting in a more powerful contextual representation.

3) Attention Mechanism in Information Exchange: In order to verify the mutual channel attention mechanism does make the information exchange network learn useful semantic information, we compare the estimation of the information exchange network with that of network which increases the mutual attention mechanism. The gradient constraints and L2 loss are used as the loss function to train the network. We use the first channel configuration in Table.II. The results are illustrated in Table.IV.

It can be seen from the comparison that the network prediction capability after increasing the channel attention module is stronger than the network without increasing the attention mechanism. Automatically regulation of semantic information by the network itself is a potential way. We also use the self-attention mechanism [42], which is called as "SCA" in the second row of the Table.IV. However, increasing the mutual channel attention mechanism significantly improves the prediction ability of the network. The meaningful semantic information of additional feature groups is boosted with high weights while semantic redundancies are weakened with small weights adaptively.

4) The Influence of Loss Function: In order to prove that our loss function can further improve the depth estimation, we use different loss functions to train the network. We use the fourth channel configuration in Table.II and abandon mutual attention module. The test results are as shown in Table.V. It can be seen that the log-cosh loss of log space can get better effect than L2 loss and Berhu loss. Thanks to piecewise property, log-cosh loss shows a good balance when the residual is varying. The gradient constraints can improve the overall performance of the network. It makes the depth gradient consistent with the ground-truth in high order.

C. State-of-the-Art Comparisons

In this section, we compare our approach with recent popular depth estimation methods on NYU-Depth V2 and KITTI datasets. We apply our information exchange network with mutual attention module and regress the depth values with log spatial log-cosh loss and gradient constraints. The results are reported in Table.VI and Table.VII respectively.

1) Compared Methods: In Table.VI, the first row is the results in [24] which apply an up-sampling scheme. The second row is the results of deep convolutional neural fields (DCNF) with fully convolutional network and super-pixel pooling in [28]. The third row is the results of neural regression forest (NRF) in [33]. The fourth row is the results in [34] which generate features and refine prediction with a single multi-scale convolutional architecture. The fifth row is the results in [16] which apply an up-sampling scheme. The sixth row is the results in [32] which estimate depth as classification combining conditional random fields. The seventh row is

TABLE VI

DEPTH ESTIMATION RESULTS ON NYU-DEPTH V2 TEST DATASET

	Accuracy			Error		
	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$	REL	LOG10	RMSE
Wang et al.[24]	0.605	0.890	0.970	0.21	0.094	0.745
Liu et al.[28]	0.650	0.906	0.976	0.213	0.087	0.759
Anirban et al.[33]	–	–	–	0.187	0.078	0.744
Eigen et al.[34]	0.769	0.950	0.988	0.158	–	0.641
Laina et al.[16]	0.811	0.953	0.988	0.127	**0.055**	0.573
Cao et al.[32]	0.819	0.965	0.992	0.141	0.060	0.540
Xu et al.[30]	0.806	0.952	0.986	0.125	0.057	0.593
Zhang et al.[46]	**0.835**	0.962	0.992	0.144	–	**0.501**
Kim et al.[47]	0.825	**0.976**	**0.993**	**0.117**	–	0.525
Fu et al.[17]	**0.828**	0.965	0.992	**0.115**	**0.051**	0.509
Ours	0.8258	**0.9670**	**0.9949**	0.1366	0.0583	**0.4983**

TABLE VII

DEPTH ESTIMATION RESULTS ON KITTI TEST DATASET

	Accuracy			Error		
	$\delta_1 < 1.25$	$\delta_2 < 1.25^2$	$\delta_3 < 1.25^3$	REL	RMSE	RMSELOG
			Cap 80 meters			
Liu et al.[28]	0.656	0.881	0.958	0.217	7.046	–
Eigen et al.[34]	0.692	0.899	0.967	0.190	7.156	0.270
Cao et al.[32]	0.887	0.963	0.982	0.115	4.712	0.198
Godard et al.[38]	0.836	0.935	0.968	0.136	5.763	0.236
Kuznietsov et al.[36]	0.862	0.960	0.986	**0.113**	4.621	0.189
Xu et al.[30]	0.818	0.954	0.985	0.122	4.677	–
Zhang et al.[46]	0.864	0.966	**0.989**	0.136	**4.082**	**0.164**
Kim et al.[47]	–	–	–	0.177	–	0.254
Fu et al.[17]	**0.935**	**0.984**	**0.994**	**0.072**	**2.727**	**0.120**
Ours	**0.8941**	**0.9710**	0.9843	0.1170	4.2510	0.1744
			Cap 50 meters			
Cao et al.[32]	0.898	0.966	0.984	0.107	3.605	0.187
Godard et al.[38]	0.858	0.947	0.974	0.118	4.941	0.215
Kuznietsov et al.[36]	0.875	0.964	0.988	0.108	3.518	0.179
Fu et al.[17]	**0.936**	**0.985**	**0.995**	**0.071**	**2.271**	**0.116**
Ours	**0.9004**	**0.9760**	**0.9892**	**0.1068**	**3.4393**	**0.1718**

the results in [30] which fuse representations derived from multiple output layers of CNN based on continuous CRFs and use structured attention model to regulate the amount of information transferred between corresponding features at different scales. The eighth row is the results in [46] which use progressive hard-mining network framework to preserve the cross-border details of depth maps. The ninth row is the results in [47] which integrate heterogeneous predictions from global and local networks with a deep variational model. In Table.VII, the result in [38] which exploit epipolar geometry constraints to perform single image depth estimation with an unsupervised manner and the result in [36] which predict depth map in a semi-supervised way are also reported.

2) NYU-Depth V2 Results: For the indoor dataset NYU-Depth V2, as we can see from the table, our information exchange network achieves competitive performance compared with other methods excepting ordinal regression [17]. We guess this is related to that we ignore the order information between depth values. This is the point we will improve. Compared to the results of [16], [24], our information exchange network achieves better performance and we can maintain high-resolution prediction instead of up-sampling final prediction. Compared to the results of [34], [47], our multi-scale contextual information exchanging is more effective. Compared to [30], our mutual attention focuses on adjusting the role of semantic information in depth estimation. We also show some qualitative results in Figure.6, from which we can see our method yields better visualizations in general. Our estimations perform well in finding the semantic boundary between the objects.

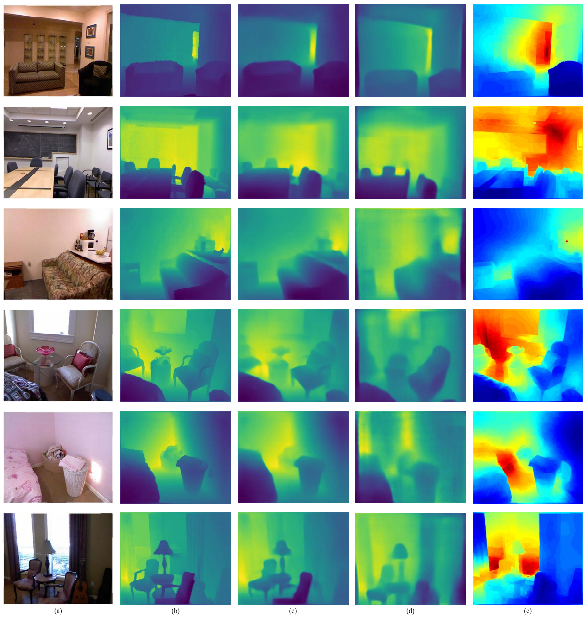


Fig. 6. Visualizations of depth prediction on NYU-Depth V2. (a) RGB images. (b) Ground truths. (c) Our estimations. All colormaps are scaled equally for better comparison. (d) Results of Eigen *et al.* [34]. (e) Results of Liu *et al.* [28]. Yellow or red color indicates depth is high, blue color indicates depth is low.

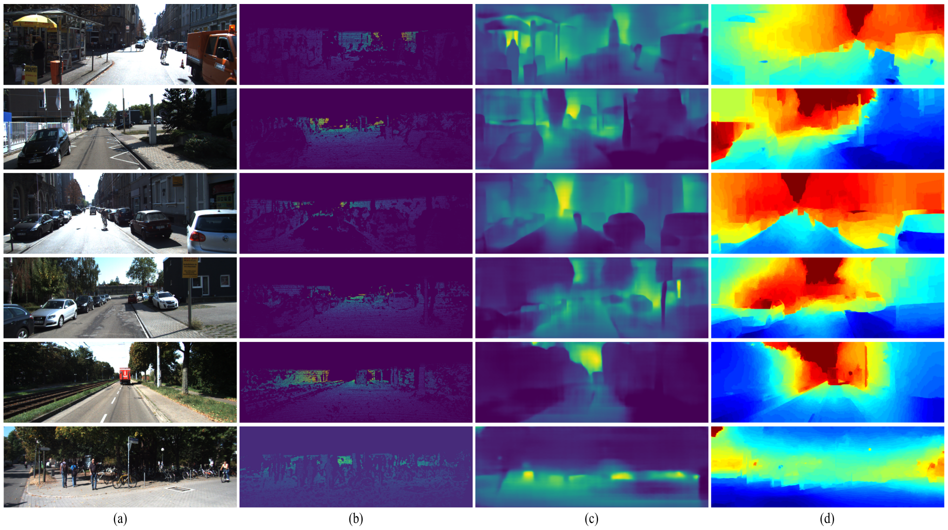


Fig. 7. Visualizations of depth prediction on KITTI. (a) RGB images. (b) Ground truths. (c) Our estimations. (d) Results of Liu *et al.* [28].

3) KITTI Results: For the street view dataset KITTI, the results are reported in Table.VII. We train our model on the full depth range, and test it on data with different depth ranging from 0m to 80m and 0m to 50m. The last row is our depth estimation result. As we can see from the table, our approach also has powerful ability to estimate depth outdoor. We also show some qualitative results in Figure.7. It is noted the ground truth is not in line with peoples perception of depth in fact because the depth values are sampled by Velodyne HDL-64E rotating 3D laser scanner. The data in the top of images is missing due to the sensor height limit. However, we estimate depth with regression from original image so that our results are more in line with the actual depth situation. Besides, the proposed method predicts more accurate depth values. For instance, the large-depth (yellow) regions in these examples, and the contours of objects. The first row and the second row in the figure show that our approach is good at constructing the depth of areas that lack illuminance. We think it benefits from the employments of diverse contextual information. With the introduce of high-level semantic information, our estimation can distinguish small objects, such as road signs, street facilities and pedestrian. The depth of road can also be predicted to be smooth and not simply affected by color or shadow in the last two row of Figure.7. The last row gives a failure case. We think it is mainly because of the complexity of scene in where the objects are too thin and small.

Considering the unavailability of depth annotation in larger automotive datasets, we test the trained model of KITTI explicitly on nuScenes and Apolloscapes. But the estimations are unsatisfactory. We think the main reason is that our model is learning based. Without training on specific dataset, the estimations of such complex scenes will be unsatisfied.

V. Conclusion

In this paper, we address the monocular depth estimation from a single image as a regression of information exchange network with mutual attention module. Our method is motivated by three aspects: (I) To obtain high-resolution depth map, depth estimation networks require incorporating multi-scale context; (II) Semantic information play an important role in monocular depth estimation; (III) Depth space and MSE has inherent difficulty in solving depth estimation. To this end, we first introduce a simple depth estimation network which takes advantage of aggregating multiple receptive fields to directly obtain a high-resolution multi-scale contextual feature map. Second, we aggregate useful sematic information by mutual attention module. Last, an effective special log-cosh loss and gradient constraints are intergraded to improve the training of our network. The proposed method achieves competitive performance on the NYU-Depth V2 and KITTI datasets.

References

[1] D. Hoiem, A. A. Efros, and M. Hebert, “Closing the loop in scene interpretation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[2] S. Gao, Z. Han, C. Li, Q. Ye, and J. Jiao, “Real-time multipedestrian tracking in traffic scenes via an RGB-D-based layered graph model,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2814–2825, Oct. 2015.

[3] D. Hoiem, A. A. Efros, and M. Hebert, “Automatic photo pop-up,” *ACM Trans. Graph.*, vol. 24, no. 3, pp. 577–584, Jul. 2005.

[4] T. Roddick, A. Kendall, and R. Cipolla, “Orthographic feature transform for monocular 3D object detection,” 2018, *arXiv:1811.08188*. [Online]. Available: http://arxiv.org/abs/1811.08188

[5] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, “Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation,” *Int. J. Comput. Vis.*, vol. 112, no. 2, pp. 133–149, Apr. 2015.

[6] A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg, “Joint semantic segmentation and 3D reconstruction from monocular video,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 703–718.

[7] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, “The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 103–110.

[8] K. Park, S. Kim, and K. Sohn, “High-precision depth estimation using uncalibrated LiDAR and stereo fusion,” *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 321–335, Jan. 2020.

[9] H. Ha, S. Im, J. Park, H.-G. Jeon, and I. S. Kweon, “High-quality depth from uncalibrated small motion clip,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5413–5421.

[10] A. Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009.

[11] D. Hoiem, A. A. Efros, and M. Hebert, “Recovering surface layout from an image,” *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, Jul. 2007.

[12] A. Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1161–1168.

[13] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

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

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

[18] B.-S. Kim, P. Kohli, and S. Savarese, “3D scene understanding by Voxel-CRF,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1425–1432.

[19] A. Gupta, A. A. Efros, and M. Hebert, “Blocks world revisited: Image understanding using qualitative geometry and mechanics,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 482–496.

[20] X. Liu, Y. Zhao, and S.-C. Zhu, “Single-view 3D scene parsing by attributed grammar,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 684–691.

[21] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.

[22] A. Saxena, S. H. Chung, and A. Y. Ng, “3-D depth reconstruction from a single still image,” *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 53–69, Dec. 2007.

[23] M. Liu, M. Salzmann, and X. He, “Discrete-continuous depth estimation from a single image,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 716–723.

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

[25] L. Ladicky, J. Shi, and M. Pollefeys, “Pulling things out of perspective,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 89–96.

[26] B. Liu, S. Gould, and D. Koller, “Single image depth estimation from predicted semantic labels,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1253–1260.

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

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

[29] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, “Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5354–5362.

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

[31] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, “Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1119–1127.

[32] Y. Cao, Z. Wu, and C. Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[33] A. Roy and S. Todorovic, “Monocular depth estimation using neural regression forest,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5506–5514.

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

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

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

[37] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 740–756.

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

[39] R. Ji *et al.*, “Semi-supervised adversarial monocular depth estimation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 20, 2019, doi: 10.1109/TPAMI.2019.2936024.

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

[41] C. Colombo, A. Del Bimbo, and P. Pala, “Semantics in visual information retrieval,” *IEEE MultimediaMag.*, vol. 6, no. 3, pp. 38–53, Jul./Sep. 1999.

[42] J. Fu *et al.*, “Dual attention network for scene segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[43] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 746–760.

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

[45] A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5188–5196.

[46] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, “Progressive hard-mining network for monocular depth estimation,” *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3691–3702, Aug. 2018.

[47] Y. Kim, H. Jung, D. Min, and K. Sohn, “Deep monocular depth estimation via integration of global and local predictions,” *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, Aug. 2018.

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

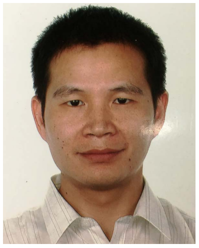
[49] J. Jiao, Y. Cao, Y. Song, and R. Lau, “Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 53–69.

[50] P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano, “Geometry meets semantics for semi-supervised monocular depth estimation,” in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 298–313.

Wen Su was born in 1992. She received the B.E. degree in engineering from the Automation Department, University of Science and Technology of China, in 2013, and the Ph.D. degree in control science and engineering from the University of Science and Technology of China in 2018. She is currently working with the Virtual Reality Laboratory, Zhejiang Sci-Tech University. Her current research interests include image segmentation and depth scene understanding based on deep learning.

Haifeng Zhang was born in 1993. He received the B.E. degree from the China University of Geosciences, Beijing, in 2015. In 2015, he was sent to the Automation Department, University of Science and Technology of China, for Ph.D. degree. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China. His research interests include face recognition and facial expression recognition.

Quan Zhou (Member, IEEE) received the B.S. degree in electronics and information engineering from the China University of Geosciences, Wuhan, China, in 1998, and the M.S. and Ph.D. degrees in communication and information system from the Huazhong University of Science and Technology, China, in 2006 and 2013, respectively. He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition. He has published many articles, including the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), PR, ICIP, ICASSP, ACCV, and ICPR. He serves as a TPC Member of many international conferences and a Reviewer for a series of SCI journals, including the IEEE TIP, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, PR, and *Neurocomputing*.



Wenzhen Yang received the M.S. and Ph.D. degrees from Zhejiang University in 2001 and 2007, respectively. He is currently working as a Professor and the Chair of the Virtual Reality Laboratory, Zhejiang Sci-Tech University. He has published more than 40 technical publications and proceedings. His research interests include VR, HCI, and AI.

Zengfu Wang (Member, IEEE) was born in Hefei, Anhui, China, in 1960. He received the B.S. degree from the University of Science and Technology of China in 1982 and the Ph.D. degree in control engineering from Osaka University, Japan, in 1992. He is currently a Professor with the Institute of Intelligent Machines, Chinese Academy of Sciences, and the Department of Automation, University of Science and Technology of China. He has published more than 300 journal articles and conference papers. His research interests include computer vision, human–computer interaction, and intelligent robots. He received the Best Paper Award at the 17th ACM International Conference on Multimedia 2009 and the IET Image Processing Premium Award 2017.