Attention-based Context Aggregation Network for Monocular Depth Estimation

Yuru Chen, Haitao Zhao, Zhengwei Hu

1 首先将深度估计作为多分类问题 然后提出软序数推断出连续深度值
2 ACAN融合图像级和像素级的内容
3 为了解决深度图和RGB不一致的问题，提出注意力loss减小信息熵

School of Information Science and Engineering, East China University of Science and Technology, China

***Abstract*—Depth estimation is a traditional computer vision task, which plays a crucial role in understanding 3D scene geometry. Recently, deep-convolutional-neural-networks based methods have achieved promising results in the monocular depth estimation field. Specifically, the framework that combines the multi-scale features extracted by the dilated convolution based block (atrous spatial pyramid pooling, ASPP) has gained the significant improvement in the dense labeling task. However, the discretized and predefined dilation rates cannot capture the continuous context information that differs in diverse scenes and easily introduce the grid artifacts in depth estimation. In this paper, we propose an attention-based context aggregation network (ACAN) to tackle these difficulties. Based on the self-attention model, ACAN adaptively learns the task-specific similarities between pixels to model the context information. First, we recast the monocular depth estimation as a dense labeling multi-class classification problem. Then we propose a soft ordinal inference to transform the predicted probabilities to continuous depth values, which can reduce the discretization error (about 1% decrease in RMSE). Second, the proposed ACAN aggregates both the image-level and pixel-level context information for depth estimation, where the former expresses the statistical characteristic of the whole image and the latter extracts the long-range spatial dependencies for each pixel. Third, for further reducing the inconsistency between the RGB image and depth map, we construct an attention loss to minimize their information entropy. We evaluate on public monocular depth-estimation benchmark datasets (including NYU Depth V2, KITTI). The experiments demonstrate the superiority of our proposed ACAN and achieve the competitive results with the state of the arts. The source code of ACAN can be found in https://github.com/miraiaroha/ACAN**

ASPP结构提取多尺度信息 但是预定义的扩张率无法获得不同场景的连续内容信息并引入网格伪影 故提出注意力导向的融合网络ACAN

I. Introduction

Depth information has a significant impact on understanding 3D scenes and can benefit the tasks such as 3D reconstruction[36] , 3D object detection [37], visual simultaneous localization and mapping (SLAM) [38], [18], and autonomous driving [9]. Estimating the pixel-wise depth of scenes from RGB images has triggered wide research recently in the computer vision community. The goal of depth estimation is to assign each pixel in an image the distance between the observer and the scene point represented by this pixel. Estimating the depth from a single monocular image is ill-posed without any geometric cues or priors. Therefore the previous works mainly focus on the stereo vision [13], [29], in which the binocular images or multi-view images are adopted to obtain the disparity map, and the depth information can be further reconstructed from the disparity map by utilizing the camera parameters. However, the drawbacks of stereo matching lie in the blind areas of the prediction due to the existence of occlusion, and the predicted results might be distorted by inaccurate camera parameters.

Recently, deep-neural-network-based methods have been widely used in computer vision tasks and achieved great performances. Convolutional neural networks (CNNs) have been proved effective for image classification. Simultaneously, people have applied CNN to dense labeling tasks, such as monocular depth estimation [6], [5], semantic segmentation [2], [49] and edge detection [43] by modifying the network structure of CNN.

Despite the above success, there still has existed some key challenges in monocular depth estimation tasks. In common deep-CNN-based image-processing, the spatial scales of feature maps continue to shrink as the network goes deeper due to the successive pooling and stride operations, which allows the deep CNN to learn the increasingly abstract representations and fuse the global features to obtain the image-level prediction. However, this translation invariance property may hinder the dense prediction tasks, such as semantic segmentation and depth estimation, where detailed spatial information and image structure are crucial. To overcome this problem, some previous works utilize the skip connection[30] to combine the feature maps produced by shallow layers and deep layers of the same spatial scales. Moreover, the intermediate supervision [27], [42] is applied to the multi-scale cues to progressively refine the prediction. In other works [47], [46], the application of dilated convolution maintains the resolution while extending the receptive fields and without introducing extra parameters.

Another challenge comes from the depth distribution of objects in the scene. Huang et al. [15] studied the statistics of range images of natural scenes (called depth maps in the depth estimation field), which showed that the range images can be decomposed into piecewise smooth regions that show little dependencies with each other and the sharp discontinuities typically exist in the object boundaries. Therefore, the concept of "objects" in the scene can be better defined in terms of changes in depth rather than some low-level features, such as color, intensity, texture, lighting etc. From this perspective, depth estimation as a classification task can be regarded as a generalized semantic segmentation task while the labels between pixels are not independent. Accordingly, the key point in depth estimation is how to capture the long-range context information of intra-object and inter-object. Yu et al. [46] used serialized layers with increasing dilation rates to extend the receptive fields of convolutional kernels, while the research works [2], [3] implement an "atrous spatial pyramid pooling

(ASPP)" framework to capture multi-scale objects and context information by placing multiple dilated convolution layers in parallel. However, the discretized and limited dilation rates cannot cover the complex and continuous scene depth and easily introduce the grid artifacts [40], which can be found in Fig. 9.

In view of the challenges, this paper proposes a novel depth estimation algorithm, called the attention-based context aggregation network (ACAN), to tackle the monocular depth estimation problem. The deep residual architecture [11] is adopted by ACAN, where dilated convolutions are used to maintain the spatial scale. To extract the continuous pixel-level context information, the self-attention module [39], [41], [25] is plugged into our model to approximate depth distribution of scenes by learning an attention map that carries the normalized similarities between all the pixels. According to the learned attention map, we can obtain the context information of each pixel. Different from prefixed or prestructured local kernels, our proposed attention model can obtain adaptive similarities, which reflect the relationships between each pixel and any other pixels in the whole feature map. Instead of using predefined regions and extracting sparse context information in ASPP, the proposed ACAN can learn the attention weights associated with meaningful contextual areas, resulting in predicting the piecewise smooth depth. The comparison between ASPP and our proposed ACAN can be seen in Fig. 2. To reduce the inconsistency between RGB image and depth map, KL divergence is adopted to model the divergence between the distribution produced by the self-attention model and the distribution constructed by the corresponding ground truth depth. To further incorporate the image-level information for depth estimation, the image-pooling [3], [25] is utilized in this paper. Finally, our proposed soft ordinal inference translates the predicted probabilities into the continuous depth values and produce more realistic transitional regions.

The main contributions of this paper can be summarized as follows:

- We propose a pixel-level attention model for the monocular depth estimation that can capture the context information associated with each pixel. In addition, the aggregation of pixel-level context and image-level context is effective to promote the estimation performance. Our experimental results demonstrate that the proposed pixel-level attention model outperform the ASPP based model since the generated pixel-level context information of ACAN is flexible and continuous, and therefore avoid the grid effect.
- To eliminate the large semantic gap from 2D image texture and depth map, we introduce KL divergence as our attention loss to minimize the divergence between the distribution of the attention map and the distribution of the similarity map constructed by the ground truth depth. The effectiveness of the attention loss is confirmed by our ablation experiments.
- An easy-implemented soft inference strategy is proposed in this paper, which can reduce the discretization error and produce more realistic depth map compared with the naïve hard inference.

II. Related Work

Estimating the depth of a scene is a traditional task in computer vision and has been studied for a long period. As a pioneering work, Saxena et al. [34] infer the depth from monocular cues based on Markov Random Field (MRF), and further develop their method in [33], where the smoothness assumption is imposed to the superpixels to enforce the neighboring constraint. Their work later extended for the 3D model generation [35]. In [22], semantic labels are incorporated into the MRF framework to guide the depth estimation. Ladicky et al. [17] showed that the property of perspective geometry could be used to learn a much simpler classifier to predict the likelihood of a pixel instead of a pixel-wise depth classifier. All these works provide novel thoughts, while most of them rely on strong geometric constraints and hand-crafted features thus limit their models to generalize to diverse scenarios.

Recently, a large body of works adopts the deep neural network for monocular depth estimation[6], [5], [14], [3], [18], [45]. The seminal work of Eigen et al. [6] first proposed a multi-scale coarse-to-fine model, where the fine network refines the global prediction from coarse network to produce a more detailed result, and the innovative scale-invariant loss is proved an effective loss function both for training and evaluation. They then extended their model to a three-scale architecture for three dense labeling tasks, i.e. predicting normal, label and depth [5]. In order to solve the heavy-tailed effect of depth values reported in [31], Laina et at. [18] presented that the reverse Huber loss [53] is more appropriate than standard L2 regression loss for depth estimation since Huber loss is more sensitive to small errors. While the deep CNN-based methods are excellent at extracting image features, they are weak in reconstructing high-resolution images due to the down-sampling operation and lack of structural constraints, therefore, often obtain the depth estimation with distorted boundaries and counterfeit regions. To tackle this problem, Hu et al. [14] proposed the notable loss function whose three items are complementary with each other and the loss function is edge-aware. Garg et al [8] proposed an unsupervised framework for single view depth estimation with a photometric reconstruction loss between stereo pairs. Under this setting, Godard et al. [10] further proposed a combination of an L1 loss and the structural similarity index (SSIM) term [52] as the reconstruction loss and explicitly imposed a spatial smoothness constraint [12] for the synthesized image. Chen et al. [4] regarded the depth estimation as an image-to-image translation task, additionally utilized an adversarial loss with the discriminator as a structural penalty.

Besides the above methods using the task-specific loss or geometric prior to supervise the network learning, there exists another research route that fuses multi-scale information in CNNs for pixel-level prediction [43], [30], [47], [46], [3], [16]. Most of them applied an encoder-decoder architecture, where a reliable encoder adaptively learns the hierarchical features of input RGB images. In the decoder, the specially designed building blocks are employed to recover the spatial resolution or leverage the multi-scale context to restore the finer details. Laina [18] introduced an up-sampling block to

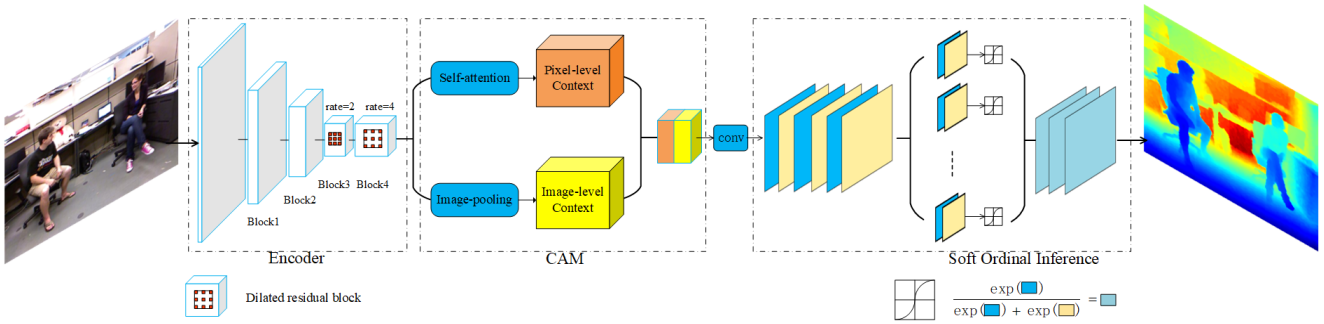


Figure 1: Network architecture. The ResNet is adopted by the ACAN as the encoder, where the cascaded 2-dilated and 4-dilated convolutions are used to avoid the over-downsampling. In the decoder, the CAM is proposed to extract and aggregate both the pixel-level and image-level context. Finally, our proposed soft ordinal inference will translate the predicted probabilities into continuous depth values. 级联的2和4倍扩张率卷积避免过度下采样，decoder中CAM提取像素级和图象级特征，最后是软序关系推断

improve the output resolution. In the research works [44], [24], [20], [23], [45], the conditional random field (CRF) based models have been utilized for the multi-scale features to estimate the fine-grained depth maps. Kim et al. [16] proposed a deep variational model that integrates the predictions from the global and local networks. In the research works [30], [10], [48], [19], skip connections were added to concatenate the detail-abundant features from the encoder with the decoder features of the corresponding scales. Although these works give the impressively sharp inferences, they also introduce the inevitable artifacts in some highly textured regions [48], [26]. To address this problem, Fu et al. [7] employed the dilated convolutions to capture context information in multiple scales, a typical example is ASPP [3], which has been well studied in semantic segmentation [2], [46], [3]. While the dilated-convolution-based methods have achieved the state-of-the-art, the dilated kernels introduce a sparse sub-sampling of activations, which results in an inherent problem identified as "gridding" [3].

To deal with the gridding problem, different from using prefixed structures of the dilated kernels, we design an attention model to extract the continuous multi-scale context by adaptively learning the pixel-level similarity map. The output features of the decoder can be computed by a weighted sum of contextual regions, which is essential for the fine-grained depth estimation. Moreover, due to our designed attention loss, the ambiguity caused by the large semantic gap could be partly eliminated and the produced attention map could be task-specific. CRF is widely adopted to obtain the pixel-level pairwise similarities as the context information [2], [44], [50], [21], [20], [23]. However, the similarities are patch-wise and only able to compute between the pixel and its local neighborhoods. Armed with the pixel-level attention, which could be regarded as a global structural extractor, our proposed ACAN can capture the long-range dependencies of intra-object by directly computing interactions between widely scattered pixels which share similar depth values.

Although estimating a depth range is more robust than estimating a depth value for each pixel and the classification strategy can put different weights on different depth ranges according to the tasks of depth estimation [1], the naïve hard-threshold-based depth inference ignores the predicted confidence of the depth distribution and usually introduces the stepped artifacts [7], [1]. In this paper, taking the full advantage of the output confidence of the proposed network, we propose a soft inference strategy to reduce the discretization error and eliminate the stepped artifacts.

III. Methods

This section introduces the architecture of our proposed ACAN and associated loss functions for the monocular depth estimation, which maps an RGB image to its corresponding depth map in an end-to-end fashion.

A. Network Architecture

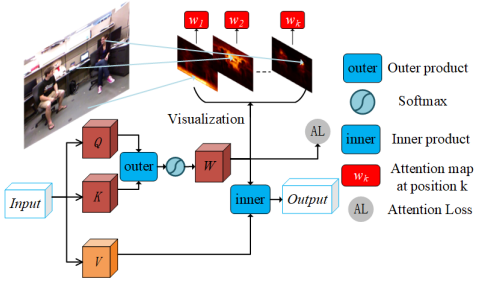
The network architecture is illustrated in Fig. 1, which also uses the encoder-decoder framework. We consider the ResNet as the encoder to extract the dense features of RGB image. ResNet shows great gradient-propagating capability in deeper networks by adding identity branches to plain network, which is essential for depth estimation due to its large receptive field [18]. However, the over-downsampling of original ResNet may hinder the reconstruction of the fine-grained depth map. Instead, we replace the block3 and block4 in ResNet with 2-dilated and 4-dilated residual blocks, which favor for the initialization of pre-trained parameters and maintain the scale of the subsequent feature maps [3], [46]. 过下采样会阻碍精细的深度预测

In the decoder, we propose a novel building block that called the context aggregation module (CAM) to enable the network to capture the discriminative image-level and pixel-level context information. We finally jointly train our model using the combination of attention loss and ordinal loss. We then describe CAM and the training losses in detail.

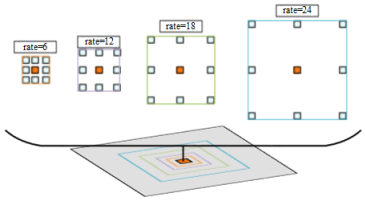
B. Context Aggregation Module

As illustrated in Fig. 1, the CAM includes two branches. The top branch is a pixel-level attention model, i.e. self-attention, the bottom branch is the image pooling operation. In the end, the resulting output features from the two branches are concatenated and passed to the subsequent classifier.

研究该模块与Non-local网络的区别



(a) self-attention module



(b) ASPP

Figure 2: (a) Our pixel-level attention model can capture the global and dense context for each location, while ASPP only parallelizes a limited amount of convolution kernels thus resulting in the sparse sampling. In addition, the attention loss is proposed to reduce the semantic gap between RGB image and depth map.

ASPP有限的卷积核导致稀疏采样，注意力loss可以减少RGB和深度图之间的语义差距

Self-Attention: The self-attention module maps a query and a set of key-value pairs to an output, where query, key and value represent three feature vectors extracted by the input via three transformation functions respectively. The output is computed as a weighted sum of the values in the feature space, where the weight assigned to each value is computed by a pairwise function of the query with the corresponding key. The details of the self-attention model are illustrated in Fig. 2a.

Specifically, as demonstrated in Fig. 2a, the feature map $x \in \mathbb{R}^{N \times C_{in}}$ inputted to the self-attention module is first encoded into two embedded features, i.e. key feature $K \in \mathbb{R}^{N \times C_K}$ and query feature $Q \in \mathbb{R}^{N \times C_Q}$, where $K = \phi(x)$, $Q = \varphi(x)$, ϕ and φ are the transformation functions, N is the number of spatial positions, i.e. $N = H \times W$, and $C_K = C_Q < C_{in}$. The normalized attention weight $w_{i,j}$ can be computed by a pairwise function $\mathcal{F}$ as follows,

$$w_{i,j} = \frac{1}{\Omega_i(x)} \mathcal{F}(x_i, x_j) \quad i, j = 1, \ldots, N \tag{1}$$

xi，xj分别来自Q，K

Where $\Omega_i(x)$ is the normalized factor, defined as $\Omega_i(x) = \Sigma_{j=1}^{N} \mathcal{F}(x_i, x_j)$. To be more specific, we consider the embedded Gaussian as the pairwise function.

$$\mathcal{F}(x_i, x_j) = e^{\frac{\phi(x_i)^T \varphi(x_j)}{\sqrt{C_K}}} \tag{2}$$

where $\sqrt{C_K}$ is the scaling factor to prevent values produced by outer-product operation growing large in magnitude, thus pushing the attention weights into saturation. Therefore, the pixel-level attention map $W \in \mathbb{R}^{N \times N}$ can be denoted equivalently as

$$W = \text{softmax}\left(Q^T K\right) \quad \text{内积？} \tag{3}$$

By virtue of the learned attention weights, the output of self-attention module at position i can be defined as

$$c_i = \sum_{j=1}^{N} w_{i,j} \psi\left(x_j\right) \tag{4}$$

是逐点的attention

where ψ transforms x into value feature $V \in \mathbb{R}^{N \times C_V}$. By this way, the process of feature extraction is enhanced via explicitly aggregating the context representation of the i^{th} pixel according to the learned attention weights. In this paper, we choose the 1×1 convolutions followed by the batch normalization layer and ReLU activation function as ϕ and φ and they share the same parameters.

Image Pooling: The image pooling has been widely used to produce a class-specific activation map [51]. We first apply global average pooling (GAP) over the whole image to reduce the 3D input feature maps to a 1D context vector, i.e., output one response for every input feature map. Then by replicating feature vector to the size of the input feature map, we can achieve the image-level context map, which carries the mixture of information belonging to different categories (channels) and helps to clarify local confusions [25]. Essentially, we discover that the GAP is similar to the channel-wise attention mechanism, the difference is that the latter applies a softmax to the context vector produced by the GAP to get an output probability. The effectiveness of image pooling lies in its class-awareness. Given the input image of a scene, the GAP can obtain its statistic prior to the features of the whole image. Our experiment at Section IV-E3 confirms our assumption.

GAP类似于CAB，只不过没有softmax

C. Training Loss

Our overall training loss $\mathcal{L}$ includes two items

$$\mathcal{L} = \alpha_{att} L_{att} + \alpha_{ord} L_{ord} \tag{5}$$

Where L_{att} is the attention loss and L_{ord} is the ordinal loss, α_{att} and α_{ord} are the coefficients.

Attention Loss: As illustrated in Fig. 2a, to bridge the semantic gap between the RGB image and depth, we consider the KL divergence as the attention loss for training the attention model, which measures the distance between the attention weights produced by the self-attention with respect to its ground truth,

$$L_{att} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j}^{*} \ln\left(\frac{w_{i,j}^{*}}{w_{i,j}}\right) \tag{6}$$

Where $w_{i,j}$ is the attention weights produced by Eq. 1 and $w_{i,j}^{*}$ can be computed by the ground truth depth values of i^{th} pixel and j^{th} pixel as follows, 类似于Nonlocal网络

$$w_{i,j}^{*} = \frac{\exp\left(\ln d_{\max} - |\ln d_i^{*} - \ln d_j^{*}|\right)}{\sum_{j=1}^{*} \exp\left(\ln d_{\max} - |\ln d_i^{*} - \ln d_j^{*}|\right)} \tag{7}$$

where $d_{\max}$ denotes the preset value that is slightly larger than the maximum depth value in the dataset. It is noted that

$w_{i,\cdot}$ (the i_{th} row of w) is the normalized attention distribution of the i^{th} pixel. Without using L_{att}, our attention model can also produce certain plausible attention map according to the extracted features of the image. However, this is problematic on the highly textured surface as the assumption of appearance-depth correlation is violated in these regions.

Ordinal Loss: The depth estimation is regarded as a pixel-level classification problem. Due to the severe imbalance of depth data, the samples are distributed more frequently in the small depth value intervals [19]. However, since the magnitude of the error of the large depth sample is larger than that of the small depth sample, the network may over-fit the former. Hence, we discretize the ground truth depth value d_i^* in logarithmic space into K sub-intervals equally,

$$l_i^* = \lfloor \frac{\ln d_i^* - \ln d_{\min}}{\ln d_{\max} - \ln d_{\min}} \times K \rfloor \tag{8}$$

Where $l_i^* \in \{0, 1, \cdots, K-1\}$ is the quantified label of i^{th} pixel, d_i^* is the continuous depth value of i^{th} pixel. The ordered discretization thresholds $t^k \in \{0, 1, \cdots, t^{K-1}\}$ can be obtained as follows,

$$t^k = e^{\ln d_{\min} + \frac{\ln d_{\max} - \ln d_{\min}}{K-1} * k} \tag{9}$$

The ordinal loss [28], [7] is adopted in the proposed ACAN to learn our network parameters rather than the straightforward cross entropy loss, which transfers the multi-class classification problem into a series of simpler binary classification problems, each of which only decides whether the sample is larger than t^k. The ordinal loss imposes large loss on predictions that are not consistent with the sequential property of the depth labels.

Formally, assuming $Y \in \mathbb{R}^{N \times 2K}$ denotes the output (confidence map) of the network. We can compute the ordinal loss at spatial position i,

$$\begin{aligned} \theta(y_i) &= -\sum_{k=0}^{l_i^*-1} \ln \mathcal{P}_i^k - \sum_{k=l_i^*}^{K-1} \left(1 - \ln \mathcal{P}_i^k\right), \\ \mathcal{P}_i^k &= P(l_i > k) = \frac{e^{y_{i,2k+1}}}{e^{y_{i,2k}} + e^{y_{i,2k+1}}} \end{aligned} \tag{10}$$

Where l_i is the estimated label and $\mathcal{P}_i^k$ is the ordinal probability that l^i is larger than k at position i. The image-wise ordinal loss is defined as the average of $\theta(y_i)$ over all spatial positions,

$$L_{ord} = \frac{1}{N} \sum_{i=1}^{N} \theta(y_i) \tag{11}$$

Soft Ordinal Inference: Classification instead of regression for depth estimation has been well studied in previous works, which can naturally obtain the confidence of the depth distribution [7], [19], [1]. The element in the confidence map of each class only pays attention to the specific depth interval, which simplifies the network learning. However, it introduces the discretization error, which is sensitive to the number of depth intervals. In addition, the hard-threshold-based inference strategies [7], [1] ignored the obtained probability distribution which can be an important cue during evaluating and may result in the step effect in the depth map, reported in our experiment IV-E3. Instead, we generalize the naïve hard inference to a soft version, called the soft ordinal inference to solve the above problems. The soft ordinal inference takes full advantage of the confidence of predictions and shows a strong ability to classify the transitional regions of inter-object.

After obtaining the probabilities of K binary classification for each pixel, the predicted depth d_i of hard inference can be computed as,

$$\begin{aligned} d_i &= \frac{t^{l_i} + t^{l_{i+1}}}{2} \\ l_i &= \sum_{k=0}^{K-1} \eta(\mathcal{P}_i^k \geq 0.5) \end{aligned} \tag{12}$$

where $\eta(\cdot)$ is an indicator function such that $\eta(true) = 1$ and $\eta(false) = 0$. The rounding operation of hard inference ignores the probability (or confidence) predicted by the network, which may distort the predictions of transitional regions that difficult to distinguish.

However, our soft ordinal inference can transfer the predicted probabilities to continuous depth values as follows,

$$\begin{aligned} d_i &= \frac{t^{l_i} + t^{l_{i+1}}}{2} * (1 - \mathcal{D}_i) + \frac{t^{l_{i+1}} + t^{l_{i+2}}}{2} * \mathcal{D}_i \\ l_i &= \lfloor \int_i \rfloor, \mathcal{D}_i = \int_i - l_i \\ \int_i &= \sum_{k=0}^{K-1} \mathcal{P}_i^k \end{aligned} \tag{13}$$

where $\lfloor \cdot \rfloor$ means the floor operation. $\mathcal{D}_i$ is between 0 and 1, which represents the extent to which the predicted category is close to l_{i+1}. Actually, $\int_i$ is the area under the probability distribution curve, which will be discussed in Section IV-E3.

IV. Experiments

In this section, we investigate the performance of the proposed ACAN model on two publicly available monocular depth datasets, NYU v2 Depth [36], KITTI [9].

A. NYU v2 Depth

The original NYU v2 Depth dataset [36]consists of around 240k RGB-D images of 464 indoor scenes, captured by a Microsoft Kinect camera as video sequences. Following the research works [6], [18], we use the official train\test split, where 249 scenes for training and 215 for testing. For training, we sample approximately 12k unique images with a fixed sampling frequency from each training sequence and then fill in the invalid pixels of the depth map using the colorization method, which is available in the toolbox of NYU v2 dataset. The original image resolution is 480×640, we first downsample it to 288×384 using bilinear interpolation and then randomly crop to 256×352 pixels, as inputs to the network. It is noted that the output of ACAN is 1/8 of ground truth depth in scale, we upsample the output to the desired spatial dimension bilinearly. Following [6], we use the same

online data augmentation strategies to increase the diversity of samples, which include random scaling, random rotation, color, flips, and contrast. For testing, we use the official 654 images and report our scores on a predefined center cropping by Eigen [6].

B. KITTI

KITTI dataset [9] is composed of several outdoor scenes captured by LIDAR sensor and car-mounted cameras while driving. Following [6], we use the part of raw data selected from the "city", "residual" and "road" categories for training, which including around 22k images from 28 scenes, and we evaluate on 697 images selected from the other 28 scenes. The original resolution is 375×1242, and are resized to 160×512 to form the inputs. As the target depth maps projected by the point cloud are sparse, we mask them out and evaluate the loss only on valid points in both the training and testing phases.

C. Implementation Details

We implement our proposed model using the public deep learning framework Pytorch on a single Nvidia GTX1080Ti GPU. In the proposed ACAN, both ResNet-50 and ResNet-101 are the candidates for the encoder, whose parameters are pretrained on the ImageNet classification task [32]. The depth intervals are set to 80 in all of our experiment. The learning rate strategy applies a polynomial decay, which starts with the learning rate of 2e-4 and is decayed with the power of 0.9 in the encoder. Since the shallow convolution kernels are optimized well to extract the general low-level features, we set the learning rate of the newly added decoder layers to 10 times to that of the encoder layers. SGD Optimization Algorithm is used to update the parameters, where momentum and weight decay are set to 0.9 and 5e-4 respectively. We set the weights of the different loss items to α_{att} and α_{ord}=0.1. The number of epoches is set to 50 both for KITTI and NYU v2, and batch size is set to 8. We find that further increasing the iteration number can hardly improve the performance.

D. Evaluation Metrics

Following previous works [6], we evaluate our depth predictions using the following quantitative metrics:

Threshold: % of d_i s.t. $max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) = \delta < thr, thr = 1.25, 1.25^2, 1.25^3$

RMSE(linear): $\sqrt{\frac{1}{N}\sum_i ||d_i - d_i^*||^2}$

RMSE(log): $\sqrt{\frac{1}{N}\sum_i ||\ln d_i - \ln d_i^*||^2}$

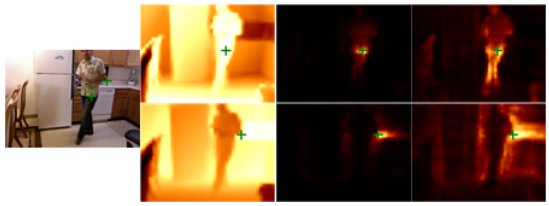
Abs Relative Difference: $\frac{1}{N}\sum_i \frac{d_i - d_i^*}{d_i^*}$

Squared Relative Difference: $\frac{1}{N}\frac{|d_i - d_i^*|^2}{d_i^*}$

Please note that N denotes the number of valid pixels.

E. Discussion on Our Work

In this subsection, we dig into the proposed context aggregation module and the training loss for the proposed ACAN.



(a) example on NYU v2

(b) example on KITTI

Figure 3: (a) example on NYU v2; (b) example on KITTI. The first column shows the RGB images from the validation set. The second column presents the ground truth contextual region computed by equation (7), the third column and fourth column present the attention map produced by our ACAN trained without and with L_att respectively. The first and second rows demonstrate the different attention maps located at "+" of the same image.

		δ_1	δ_2	δ_3	RMSE	ARE
NYU	w/o	81.9%	95.8%	98.5%	0.502	0.140
	w	82.6%	96.4%	99.0%	0.496	0.138
KITTI	w/o	91.0%	98.0%	99.4%	3.902	0.090
	w	91.9%	98.2%	99.5%	3.599	0.083
		higher is better			lower is better	

Table I: Comparisons of ACAN trained with and without attention loss

1) Effect of the attention model and attention loss: We first study the effectiveness of the proposed pixel-level attention model. For qualitative analysis, we visualize the attention maps produced by the attention model with and without attention loss respectively. Fig. 3 shows that (1) the pixel-level attention model does predict the meaningful contextual regions which capture the long-range dependencies, and is well adjusted to different scenarios adaptively. (2) The visual comparison of the produced attention maps reveals that the model trained with our L_{att} can give the more detailed and global contextual regions, which also acts as a structural extractor. For example, in the first row of Fig. 3a, the attention map without L_{att} can only capture a local contextual region at this location, while the attention map with L_{att} highlights the area of the standing man. Moreover, it also captures the context of the stair that is away from the man but similar in depth, which proves that the attention model can extract the contextual region according to the task-specific semantical correlation rather than the similarity of local and low-level features, i.e. the intensity or texture.

Quantitative results can be seen in Table I, where "w/o" denotes the model trained without the attention loss and "w" denotes the model trained with the attention loss. We can find that our ACAN with attention loss can obtain a better performance in all of the metrics.

2) Effect of the image-level feature: We conduct the ablation experiment to reveal the effectiveness of incorporating

		δ_1	δ_2	δ_3	RMSE	ARE
NYU	w/o	81.8%	96.1%	99.0%	0.504	0.140
	w	82.6%	96.4%	99.0%	0.496	0.138
KITTI	w/o	91.4%	98.2%	99.5%	3.733	0.085
	w	91.9%	98.2%	99.5%	3.599	0.083
		higher is better			lower is better	

Table II: Comparisons with and without image-level features on NYU and KITTI

(a) Input RGB images from KITTI

(b) Matrix of L2 norm

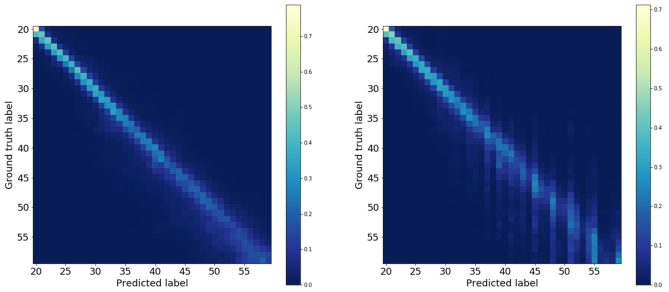
Figure 4: (a) input RGB images from KITTI, the images of the first row are from 'City' category, and the images of the second row are from 'Residential' category; (b) L2 norm of image-level context vector of the four images.

the image-level context to the proposed module. Results are shown in Table II. All of these models are built on ResNet-101. In the Table III, "w" represents an ACAN model with image-pooling block, "w/o" represents an ACAN model that sets the responses from GAP to a zero vector for a comparison.

We further explore the effect of the image-pooling module by visualizing the L2 norm matrix where each entry is calculated from the context vectors produced by the GAP between the two images. As shown in Fig. 4, the images are selected from KITTI dataset, each row in Fig.4 is from the same scenario and is visually similar. We observe that the image-pooling module has the remarkable distinguishability, as it shows significant differences between different scenes and little but not completely undifferentiated difference from the same scene. It reveals that the image-pooling module can extract the discriminative pattern of scenes.

The above experiment reveals that the image-level context information does act as a variant of channel-wise attention mechanism, which considers the class-specific statistics prior that expresses the visual characteristic of a scene. Therefore, our proposed ACAN is robust to the varied depth samples from the dataset.

3) Effect of the ordinal loss and soft ordinal inference: To demonstrate the effectiveness of the ordinal loss, we compared the depth estimation obtained by ordinal inference and that obtained using cross entropy. The experiment is evaluated on KITTI dataset. Normalized confusion matrices are plotted in Fig. 5. On the plots of confusion matrices, the columns show the predicted depth label, and the rows correspond to the true class. The diagonal elements of the plots show what percentage of the pixels the trained network correctly estimates their true classes. That is, it shows what percentage of the true and predicted labels match. The off-diagonal elements show where the depth estimation has made mistakes. From Fig. 5, it can be found that the ordinal-inference-based depth estimation achieves higher estimation accuracy.



(a) The plot of confusion matrix (ordinal inference)

(b) The plot of confusion matrix (cross entropy)

Figure 5: The plots of the normalized confusion matrices, where the predicted labels are produced by (a) ordinal inference and (b) cross entropy. Here we only show the depth labels between 20 and 60, as the samples in this range is dominant and representative in KITTI dataset.

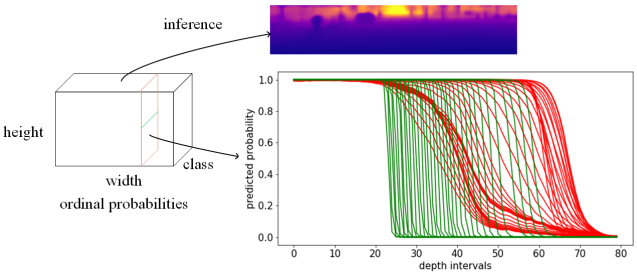


Figure 6: The typical probability distribution of the output of ordinal inference. Each curve on the plot represents the predicted ordinal probability set $\{\mathcal{P}_i^0, \mathcal{P}_i^1, \ldots, \mathcal{P}_i^K\}$ at position i.

To illustrate the effectiveness of the proposed soft ordinal inference, we give the output probabilities of the ACAN in Fig. 6, which is defined by Eq. 10. In the inference phase, the predicted depth map can be inferred from the output probability distribution. Different position has different distributions of possible depth classes, some of which are easy to estimate while others are not. For example, in Fig. 6, the depth classes of the green curves in the plot can be easily determined, while those of the red curves are hard to distinguish clearly. One plausible explanation is that the model for depth estimation has uncertainty to distinguish the depth classes from its nearby intervals at these locations. Interestingly, the probability distribution curves are roughly symmetric and centralized around the right labels. Therefore, the area under the curve of probability distribution is nearest to the ground truth depth label. Considering this statistical analysis, we propose the soft ordinal inference to estimate the depth values from the output probability effectively, which can not only infer the correct depth class but also make up the decimal error of quantization.

Furthermore, we then present both the qualitative and quantitative experimental results. All these experiments are implemented with ResNet-50. As illustrated in Table III, 'CE(hard)' represents the model trained by cross entropy and applies hard-

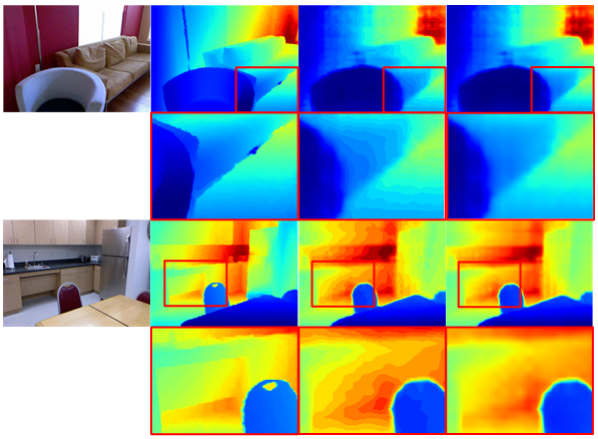


Figure 7: From left to right: input RGB image; ground truth; results of OR(hard); results of OR(soft). The images in bottom rows show the details in the red frames.

max inference, while 'CE(soft)' applies soft-weighted-sum inference proposed in [19]. 'OR(hard)' represents the model trained by ordinal loss and applies hard-threshold inference defined in Eq. 12, while 'OR(soft)' applies the proposed soft ordinal inference defined in Eq. 13. For fair comparison, the results of existing methods are also constructed with ResNet-50.

It can be found that (1) No matter using the cross entropy loss or ordinal loss, soft inference is always better than their hard counterparts. (2) The RMSEs reduce clearly (NYU v2: 1.11% for CE, 1.14% for OR; KITTI: 0.56% for CE, 1.32% for OR) while applying the soft inference. (3) The proposed soft ordinal inference achieves the best result.

The qualitative comparison of the hard inference and our soft ordinal inference are illustrated in Fig.7. As observed in Fig. 7, the results of OR(hard) produce distorted predictions on the transition area of depth while the results of OR(soft) are smooth, continuous and similar to the ground truth depth map, which well indicates that our soft ordinal inference can give the more realistic depth map without introducing the stepped artifact.

4) Comparisons with the-state-of-the-arts: 1) NYU v2 depth: We compare our proposed ACAN with state-of-the-arts on NYU v2 depth dataset. The results are shown in Table IV, and the values in Table are copied from their respective papers directly. In Table IV, the 'RX' in the brackets means the model is backboned on ResNet-X.

As we observe, our model obtains the best performance among all of the ResNet-50 based methods in all metrics and is even better than some methods built on a more stronger backbone; our ResNet-101 based model obtains competitive performance campared with some satae-of-the-arts. Specifically, in terms of RMSE, our best model outperforms the previous works in a large margin, as our quantization strategy and soft ordinal inference greatly reduce the discretization error.

Qualitative results are illustrated in Fig. 8 and Fig. 9. As we can observe in Fig. 8, the results of [18] give the semantical predictions, as their method imposed the over-downsampling to the feature maps and lack of the detail-reconstruction mechanism, resulting in their predictions corrupt into the combination of simple geometries. The results of [5] contain more details but introduce certain distortion. For example, in the third row of Fig. 8, the depth of the person is estimated inaccurate in the result of [5]. The result of [44] is blurry. In contrast, our results are detail-abundant and match the ground truth well as a whole, as our attention model can extract the global context features and is structure-aware.

Fig. 9 shows that the results of the ASPP-based method will introduce severe grid artifacts. The reason is that the kernels of the ASPP-based method are predefined elaborately, which cannot adapt to different objects in the image. However, the proposed ACAN method can produce the piecewise smoothness depth map with more details visually.

2) KITTI: Table V shows the experimental results of the proposed ACAN and the several state-of-the-art methods on KITTI dataset.

As we observe, our proposed ACAN (no matter using ResNet-50 or ResNet-101 as the encoder) achieves the excellent performance in all of the settings. Moreover, ACAN (ResNet-50) outperforms the other methods even some of their models are built on a stronger encoder. This can demonstrate that our ACAN with soft ordinal inference is a more efficient method for depth estimation.

Qualitative results are illustrated in Fig. 10. As observed in Fig. 10, the result of [6] only give the coarse and blurry predictions. The result of [10] are visually plausible, however, the depth maps of which are reconstructed indirectly via learning the disparity of the given view under the stereo constraint, which may introduce the noise. For example, the predictions of the car and the tree are confused with the background in the result of [10]. In contrast, the predictions of our method are visually satisfactory, where objects of different scales can be recognized and our model can predict the sharp boundaries as our attention model can capture the variable pixel-level context adaptively.

V. Conclusion

In this paper, we propose a deep-CNN-based method, called the attention-based context aggregation network (ACAN), for monocular depth estimation. By utilizing the self-attention model, the proposed ACAN is able to capture the long-range contextual information by learning the pixel-level attention map adaptively, which is essential for the fine-grained depth estimation. The image pooling module is also incorporated in the ACAN, which can obtain the discriminative image-level context. The aggregation of the pixel-level and image-level context is effective to promote the performance of depth estimation. Soft ordinal inference is also proposed in this paper, which takes full advantage of the output ordinal probabilities to reduce the discretization error. The experiments on NYU v2 dataset and KITTI dataset well demonstrate the superiority of our model. In the future, we plan to investigate the more effective variant of ACAN and extend our method to other dense labeling tasks, such as semantic semantation and surface normal prediction. Moreover, incorporating these tasks into the depth estimation is also our interesting work.

		δ_1	δ_2	δ_3	RMSE	RMSE(log)	ARE	SRE
NYU	Li [19]	80.8%	95.7%	98.5%	0.601	/	0.147	/
	Xu [44]	81.1%	95.4%	98.7%	0.586	/	**0.121**	/
	Laina [18]	81.1%	95.3%	98.8%	0.573	0.195	0.127	/
	CE(hard)	79.9%	95.6%	98.8%	0.536	0.188	0.151	0.118
	CE(soft)	80.0%	95.7%	98.9%	0.530	0.187	0.150	0.115
	OR(hard)	81.4%	96.0%	98.8%	0.524	0.183	0.147	0.114
	OR(soft)	**81.5%**	**96.0%**	**98.9%**	**0.518**	**0.180**	0.143	**0.110**
KITTI	Godard [10]	86.1%	94.9%	97.6%	4.935	0.206	0.190	1.515
	Zhang [48]	86.4%	96.6%	98.9%	4.082	0.164	0.139	/
	Li [19]	83.3%	95.6%	98.5%	5.325	/	0.128	/
	CE(hard)	86.9%	96.8%	99.1%	4.446	0.163	0.105	0.664
	CE(soft)	87.0%	96.9%	99.2%	4.421	0.160	0.103	0.631
	OR(hard)	91.5%	98.2%	99.5%	3.686	0.132	0.086	0.461
	OR(soft)	**91.5%**	**98.3%**	**99.5%**	**3.637**	**0.130**	**0.085**	**0.445**
		higher is better			lower is better			

Table III: Comparisons of different training losses and inference strategies On NYU and KITTI with ResNet-50

	δ_1	δ_2	δ_3	RMSE	RMSE(log)	ARE	SRE
Make3D [33]	44.7%	74.5%	89.7%	1.214	/	/	/
Ladicky [17]	54.2%	82.9%	94.1%	/	/	/	/
Liu [24]	61.4%	88.3%	97.1%	0.824	/	0.230	/
Li [20]	62.1%	88.6%	96.8%	0.821	/	0.232	/
Roy [31]	/	/	/	0.744	/	0.187	/
Liu [23]	65.0%	90.6%	97.6%	0.759	/	0.213	/
Eigen [6]	61.1%	88.7	97.1%	0.907	0.285	0.158	0.121
Eigen [5]	76.9%	95.0%	98.8%	0.641	0.214	0.158	0.121
Laina (R50) [18]	81.1%	95.3%	98.8%	0.573	0.195	0.127	/
Xu (R50) [44]	81.1%	95.4%	98.7%	0.583	/	**0.121**	/
Li (R50) [19]	80.8%	95.7	98.5%	0.601	/	0.147	/
Li (R101) [19]	82.0%	96.0%	98.9%	0.545	/	0.139	/
Yan (R101) [45]	81.3%	96.5%	99.3%	0.502	/	0.135	/
Cao (R152) [1]	81.9%	96.5%	99.2%	0.540	/	0.141	/
Li (R152) [19]	**83.2%**	96.5%	98.9%	0.540	0.187	0.134	**0.095**
Moukari (R200) [26]	83.0%	**96.6%**	**99.3%**	0.569	/	0.133	/
Our (R50)	81.5%	96.0%	98.9%	0.518	0.180	0.144	0.110
Our (R101)	82.6%	96.4%	99.0%	**0.496**	**0.174**	0.138	0.101
	higher is better			lower is better			

Table IV: Comparisons between Our Proposed Method and The Different Previous State-of-The-Arts on NYU v2 Depth Dataset

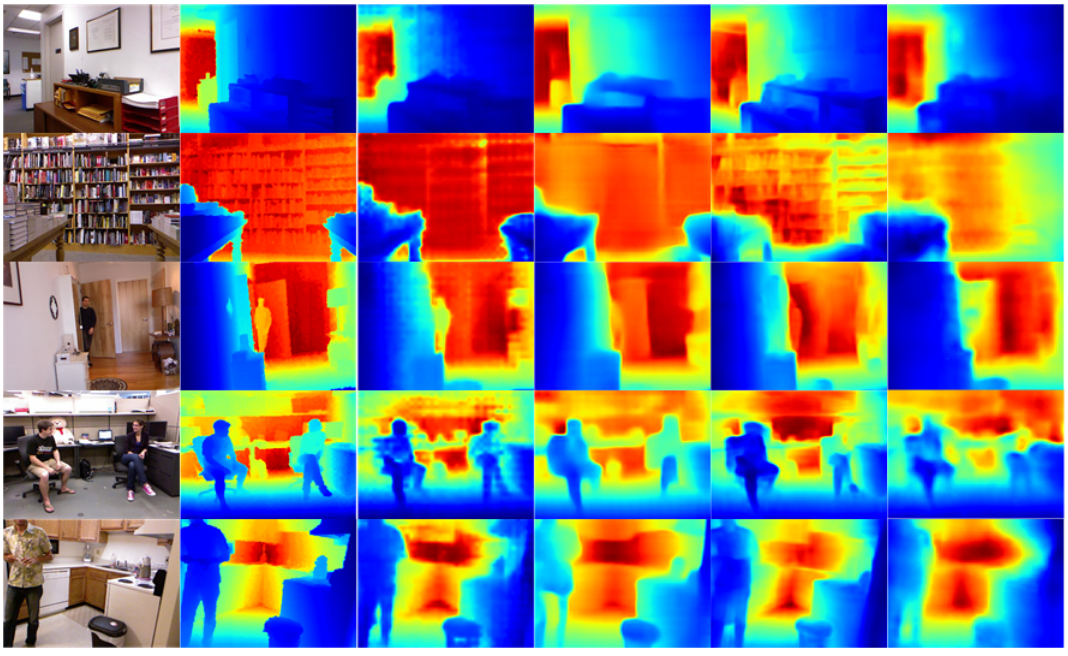


Figure 8: From left to right: input RGB images; ground truth; results of ours; results of [18]; results of [5]; results of [44].

	δ_1	δ_2	δ_3	RMSE	RMSE(log)	ARE	SRE
Liu [23]	64.7%	88.2%	96.1%	6.986	0.289	0.217	/
Eigen [6]	69.2%	89.9%	96.7%	7.156	0.270	0.190	1.515
Garg [8]	74.0%	90.4%	96.2%	5.104	0.273	0.169	1.080
Godard (R50) [10]	86.1%	94.9%	97.6%	4.935	0.206	0.114	0.898
Zhang (R50) [48]	86.4%	96.6%	98.9%	4.082	0.164	0.136	/
Li (R50) [19]	83.3%	95.6%	98.5%	5.325	/	0.128	/
Li (R101) [19]	85.7%	96.5%	98.9%	4.528	/	0.106	/
Li (R152) [19]	86.8%	96.7%	99.0%	4.513	0.164	0.104	0.697
Cao (R152) [1]	88.7%	96.3%	98.2%	4.712	0.198	0.115	/
Our (R50)	91.5%	**98.3%**	99.5%	3.637	0.130	0.085	0.445
Our (R101)	**91.9%**	98.2%	**99.5%**	**3.599**	**0.127**	**0.083**	**0.437**
higher is better				lower is better			

Table V: Comparisons between Our Proposed Method and Different State-of-The-Art Method On KITTI Dataset

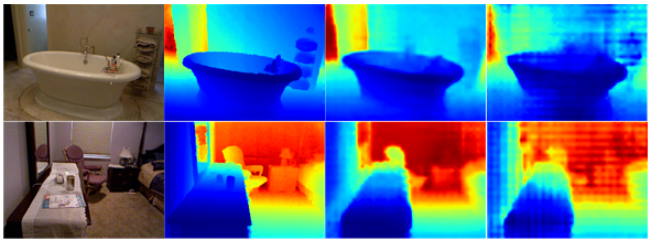


Figure 9: From left to right: input RGB images; ground truth; results of ours; results of the ASPP-based method.

REFERENCES

[1] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits & Systems for Video Technology*, PP(99):1–1, 2017.

[2] 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 Transactions on Pattern Analysis & Machine Intelligence*, 40(4):834–848, 2018.

[3] Liang Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. 2017.

[4] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018.

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. pages 2650–2658, 2014.

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *International Conference on Neural Information Processing Systems*, pages 2366–2374, 2014.

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

[8] Ravi Garg, Kumar B. G Vijay, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756, 2016.

[9] Andreas Geiger. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[10] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Computer Vision and Pattern Recognition*, pages 6602–6611, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *IEEE International Conference on Computer Vision*, pages 2360–2367, 2014.

[13] Heiko Hirschmller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 807–814, 2005.

[14] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. 2018.

[15] Jinggang Huang, Ann B. Lee, and David Mumford. Statistics of range images. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, pages 324–331 vol.1, 2000.

[16] Y Kim, H Jung, D. Min, and K Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, PP(99):1–1, 2018.

[17] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[19] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 2018.

[20] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[21] Guosheng Lin, Chunhua Shen, Ian Reid, and Anton Van Dan Hengel. Efficient piecewise training of deep structured models for semantic segmentation. pages 3194–3203, 2015.

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

[23] F. Liu, C. Shen, G. Lin, and I Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 38(10):2024–2039, 2015.

[24] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[25] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[26] Michel Moukari, Sylvaine Picard, Loic Simon, and Frédéric Jurie. Deep multi-scale architectures for monocular depth estimation. *arXiv preprint arXiv:1806.03051*, 2018.

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. pages 483–499, 2016.

[28] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] Richard Roberts, Sudipta N. Sinha, Richard Szeliski, and Drew Steedly. Structure from motion for scenes with large duplicate structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3144, 2011.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International*

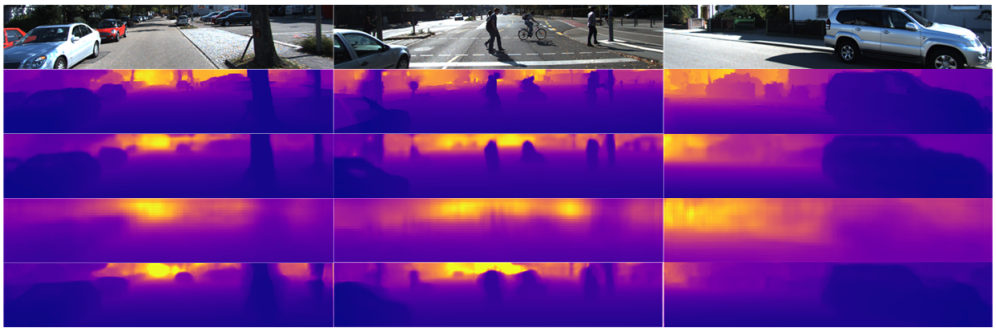


Figure 10: From up to bottom: input RGB images; ground truth; ours; results of [6]; results of [10].

Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241, 2015.

[31] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.

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

[33] A. Saxena, Min Sun, and A. Y. Ng. Learning 3-d scene structure from a single still image. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[34] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In *International Conference on Neural Information Processing Systems*, pages 1161–1168, 2005.

[35] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. 3-d depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1):53–69, 2008.

[36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. 7576(1):746–760, 2012.

[37] Martin Simon, Stefan Milz, Karl Amende, and Horst Michael Gross. Complex-yolo: Real-time 3d object detection on point clouds. 2018.

[38] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. pages 6565–6574, 2017.

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

[40] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1451–1460, 2018.

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2017.

[42] Shih En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. pages 4724–4732, 2016.

[43] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *International Journal of Computer Vision*, 125(1-3):3–18, 2015.

[44] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. pages 161–169, 2017.

[45] Han Yan, Shunli Zhang, Yu Zhang, and Li Zhang. Monocular depth estimation with guidance of surface normal map. *Neurocomputing*, 280:86–100, 2018.

[46] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[47] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. pages 636–644, 2017.

[48] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui. Progressive hard-mining network for monocular depth estimation. *IEEE Transactions on Image Processing*, PP(99):1–1, 2018.

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. pages 6230–6239, 2016.

[50] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. pages 1529–1537, 2015.

[51] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. pages 2921–2929, 2015.

[52] Wang Zhou, Bovik Alan Conrad, Sheikh Hamid Rahim, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*, 13(4):600–612, 2004.

[53] Laurent Zwald and Sophie Lambertlacroix. The berhu penalty and the grouped effect. *Statistics*, 2012.