CI-Net: Contextual Information for Joint Semantic Segmentation and Depth Estimation

Tianxiao Gao, *Student Member, IEEE*, Wu Wei*, *Member, IEEE*, Zhongbin Cai, Zhun Fan*, *Member, IEEE*, Shane Xie, *Senior Member, IEEE*, Xinmei Wang, *Member, IEEE*, Qiuda Yu

Abstract—Monocular depth estimation and semantic segmentation are two fundamental goals of scene understanding. Due to the advantages of task interaction, many works study the joint task learning algorithm. However, most existing methods fail to fully leverage the semantic labels, ignoring the provided context structures and only using them to supervise the prediction of segmentation split. In this paper, we propose a network injected with contextual information (CI-Net) to solve the problem. Specifically, we introduce self-attention block in the encoder to generate attention map. With supervision from the ground truth created by semantic labels, the network is embedded with contextual information so that it could understand the scene better, utilizing dependent features to make accurate prediction. Besides, a feature sharing module is constructed to make the task-specific features deeply fused and a consistency loss is devised to make the features mutually guided. We evaluate the proposed CI-Net on the NYU-Depth-v2 and SUN-RGBD datasets. The experimental results validate that our proposed CI-Net is competitive with the state-of-the-arts.

Index Terms—Depth estimation, Semantic segmentation, Attention mechanism, Task interaction

I. Introduction

SCENE understanding is an important yet challenging task in computer vision, which has significant contribution to visual simultaneous localization and mapping(vSLAM) system [1], robot navigation [2], augmented reality [3] and other applications. Two fundamental goals of scene understanding are monocular depth estimation [4][5][6][7] and semantic segmentation [8][9][10][11], which have been extensively researched by utilizing deep learning. Recently, some works [12][13][14] noticed the interactions between these two tasks and utilized the common characteristics to improve each other, which achieved great performance. However, most of them used a deep structure as encoder such as ResNet-101 [13], ResNet-50 [15], SE-ResNet [16], introducing a large number of downsampling and stride operations, which had negative influence [17] towards depth estimation and semantic segmentation where fine-grained information is crucial. Despite there are also some works adopting strategies like skip-connection [16], up-projection [14] and multi-scale training loss [15] to

Tianxiao Gao, Wu Wei, Zhongbin Cai and Qiuda Yu are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510000, China (e-mail: 201920116403@scut.edu.cn, eeweiwu@126.com, 201921017250@scut.edu.cn, yuqiuda@163.com)

Zhun Fan is with the College of Engineering, Shantou University, Shantou, Guangdong 515000, China (email: zfan@stu.edu.cn)

Shane Xie is with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, UK (email: S.Q.Xie@leeds.ac.uk)

Xinmei Wang is with the School of Automation, China University of Geosciences, Wuhan, Hubei 430074, China (email: wangxm@cug.edu.cn)

mitigate this problem, these schemes have great demands on computation and memory.

Another shortcoming is that current works about joint learning didn't fully exploit the contextual information of the semantic labels. As far as we know, most of them simply utilized the labels to supervise the predictions of semantic and depth splits, making limited contribution to scene understanding of the network. In [18], Chen et al. pointed out that how to obtain the correlation of inter-object and intra-object is crucial for depth estimation. Yu et al. [19] also argued such context make feature representation more robust for semantic segmentation. Therefore, could we excavate the information of labels more deeply to assist the network modeling such correlation? Moreover, most approaches achieve task interaction through adding pixel-wise features [20][16], simply sharing encoder commonly [21] or sharing parameters of convolutional layers [14]. Although these methods leveraged the correlation between tasks, their ways of fusing features are rough. For example, in [16], they fused the features via direct weighted addition and then added them to task-specific features. This simple structure may make the network difficult to learn more useful representations of the shared features.

To overcome the mentioned problems, this paper presents a network injected with contextual information (CI-Net). We adopt the dilated residual structure where the dilation operation replaces a part of downsampling layers, guaranteeing large receptive fields and avoiding introducing unnecessary parameters. To fully leverage the provided context of semantic labels, we plug a scene understanding module (SUM) with contextual supervision, which captures the similarity of pixels belong to the same classes and difference of those pertaining to different classes. Specifically, we introduce self-attention block [22][23] to generate attention map and exploit the semantic labels to create the corresponding ground truth. The attention training loss injects the contextual prior into the network, which makes the structure knowing to use correlated features for more accurate prediction. To make these two tasks deeply interacted, we presents two approaches. The first one is that we design a feature sharing module (FSM). Instead of simply adding the task-specific features, we concatenate and put them through a series of downsampling and upsampling operations for more useful representations being obtained. We also devises a consistency loss between the depth and semantic features, forcing them to maintain the intrinsic consistency of first-order relationship.

To summarize, the contributions of this paper are in three aspects:

- We propose a dilated network embedded a scene understanding module with contextual supervision to inject contextual prior about the correlation of inter-class and intra-class, predicting both the depth and semantic segmentation maps.
- We construct a feature sharing module to deeply fuse the task-specific features and put forward a consistency loss to keep the respective features consistent in the relationship with adjacent features.
- Extensive experiments are performed to demonstrate the effectiveness of our methods. And the proposed model achieves competitive results against state-of-the-art approaches of depth estimation and semantic segmentation on NYU-Depth-v2 and SUN-RGBD datasets.

II. RELATED WORKS

A. Depth Estimation from Color Images

The first research about predicting monocular depth is Saxena's work [24], which introduced Markov Random Field(MRF) to exploit the geometic priors. Later on, with the appearance of convolutional neural network(CNN), Eigen et al. [4] proposed a model intergrating the global and local information of a single image. Since then, a large number of studies using deep learning methods have emerged. Liu et al. [6] combined Convolutional Neural Fields(CRFs) with fully convolutional networks and put foward a superpixel pooling method making the inference faster. Laina et al. [25] proposed a fully convolutional residual architecture and four up-sampling models to restore the resolution of depth map. To reduce the information loss induced by excessive pooling, Fu et al. [17] employed the atrous spatial pyramid pooling (ASPP) and presented an ordinal loss to model the depth prediction as an ordinal regression problem. Inspired by [17], Chen et al. [18] proposed soft ordinal inference to exploit the predicted probabilities of the whole depth intervals and replaced ASPP with self-attention module to capture the global context. Recently, Yin et al. [26] projected the depth map to obtain the 3D point cloud, exploiting the loss between virtual normal and ground truth to train the model, which significantly improved the accuracy. There were also some works [27][28][29] who employed the geometric constraints of the consective image sequence to complete unsupervised depth prediction.

B. Depth Estimation with Semantics

Due to the common character of pixel-level prediction among different tasks, some work paid attention to studying joint learning. In [30], Wang et al. used a network to jointly predict a depth map and a semantic probability map. The network, however, is limited by only the last layer being changed for prediction. Jiao et al. [20] proposed a network with a backbone encoder and two sub-networks as decoders for respective prediction, increasing both the accuracy of depth estimation and semantic segmentation dramatically. Later, PAD-Net was proposed by Xu et al. [12], which used four intermediate auxiliary tasks, providing abundant information for prediction. In the recent research, the SOSD-Net [31] made

full use of the geometric relations between depth estimation and semantic segmentation to predict. Although these works achieved outstanding performance, they didn't exploit the feature that semantic labels could help network capture prior contextual knowledge of the scene to improve the accuracy of prediction.

C. Attention Mechanism

Xu et al. [32] was the pioneer who innovated attention mechanism into deep learning. Since then, a profusion of attention methods were designed mainly including channel attention [33], spatial attention [34] and self-attention [22]. Inspired by them, some works that incorporated attention mechanism into depth prediction have emerged. Chen et al. [35] interpolated channel attention block into the encoder and spatial block into the decoder to avoid losing structural information. Xu et al. [5] presented a fused CRF model guided by multi-scale attention. Choi et al. [16] designed an affinity propagation unit to make the depth features guided by semantic attention map.

In this work, we present a model jointly learning semantic and depth representations. We introduce a shared attention block for these two tasks with contextual supervision so that the network can understand the scene better for prediction. Moreover, we design feature sharing modules to combine the semantic and depth features, making use of the task-wise information. Notice that the similarity and discrepancy of the two kinds of features should keep consistent , we also construct a novel consistency loss.

III. METHODS

This section illustrates our proposed method for joint semantic segmentation and depth estimation from single RGB image. The first three subsections introduce the architecture of our proposed CI-Net and its sub-modules. The last subsection outlines the training losses.

A. Network Architecture

From Fig.1, it can be seen that we use the encoder-decoder scheme. For the encoder, we choose the ResNet [36] for its identity mapping tackling the vanishing gradient problem in deeper network. Another benefit of ResNet is its large receptive field [25], which is a crucial factor for depth estimation and semantic segmentation. We also adopt dilation strategy [37]. With the last two 2-dilated and 4-dilated residual blocks, the original resolution is lowered to $1/8$ instead of $1/32$, reducing the detailed information loss [38] of overdownsampling.

In the decoder, a scene understanding module (SUM) with supervised attention block is designed to fully exploit the semantic labels to obtain context prior of inter-class and intra-class, which benefits the model to understand scene better for later prediction. The network then breaks apart into two branches for estimating depth and segmenting semantics. During this stage, we present a feature sharing module (FSM) to share feature representations between these two splits. And a

consistency loss is formulated to keep the task-specific features consistent. More details about these methods are described in the following subsections.

B. Scene Understanding Module

To make full use of the superiority of joint task learning, the semantic labels are utilized for supervising the network with an attention loss to capture the correlation of the pixels belonging to same classes and the distinction of different classes. With the prior knowledge of the scene, the profitable information for prediction could be searched in a limited space instead of the whole region. Then the depth split, on the one hand, will be prevented from capturing the unrelated features. For example, the region of sky should not be used to predict the depth of ground and this behavior is hindered by the context prior for the gap between these two objects. On the other hand, the semantic one also benefits since it makes better judgments from the information of inter-class and intra-class. We encapsulate this process of obtaining contextual information in the scene understanding module, which is illustrated in the following content.

The architecture of scene understanding module is presented in Fig.2. Similar to non-local block [22], it first uses 1×1 convolutions to transform the input features $\mathbf{X} \in R^{N \times C_f}$ into query,key and value results represented by $\mathbf{Q} \in R^{N \times C_q}$, $\mathbf{K} \in R^{N \times C_k}$ and $\mathbf{V} \in R^{N \times C_v}$ respectively, where $N = H \times W$ is the resolution. Then the predicted attention map $\tilde{\mathbf{A}}$ can be obtained with

$$\tilde{\mathbf{A}} = Sigmoid(\mathbf{Q}^T \mathbf{K}), \tag{1}$$

where $Sigmoid(\cdot)$ is the sigmoid function, which ensures the attention values are in range $[0, 1]$. After that, the value results are multiplied with the attention map to capture the correlation with each pixel. By finally being added a skip connection to avoid the problem of vanishing gradient, the output $\mathbf{Y} \in R^{N \times C_f}$ can be defined as

$$\mathbf{Y} = \tilde{\mathbf{A}}\mathbf{V} + \mathbf{X} \tag{2}$$

To capture the context prior of pixels, we adopt the method of [19] to generate the ideal attention map. As it can be seen in Fig.3, given the ground truth , we can know the label of each pixel. To transform it into the relation between different pixels, the ground truth is first down-sampled into the size of $H \times W$ and then flattened into a vector $\mathbf{m}$ of size $1 \times N$. After the one-hot encoding implementation, new binary columns which indicate the presence of value from the ground truth are created, leading to a $H \times W \times C$ matrix $\mathbf{M}$, where C represents the number of total categories. The matrix $\mathbf{M}$ is then reshaped into size of $N \times C$ and finally the ideal attention map $\mathbf{A}$ is constructed with

$$\mathbf{A} = \mathbf{M}\mathbf{M}^T. \tag{3}$$

It's clear that in the ideal attention map, pixels of same classes are labeled as 1, and 0 otherwise, which aggregates the contextual information of intra-class and inter-class. And we employ the binary cross entropy loss as the attention loss:

$$\mathcal{L}_{att} = -\sum_{i,j}(\mathbf{A}_{i,j}log\tilde{\mathbf{A}}_{i,j} + ((1 - \mathbf{A}_{i,j})log(1 - \tilde{\mathbf{A}}_{i,j})), \tag{4}$$

where $A_{i,j}$ denotes the pixel at the location (i, j) of the predicted attention map.

It is worth noting that we utilize semantic labels instead of depth to inject context prior. One reason is that it is difficult to find a feasible and suitable representation of depth context. Although there exist some works proposing to use Kullback-Leibler divergence [18] or planar structures [39], their methods are limited to only depth estimation task. But for joint task learning the correlation of different objects is harmful to semantic segmentation.

C. Feature sharing Module

In the decoder, the network splits into depth and semantic branches. We design the feature sharing module (FSM) aiming to make two branches share the features with each other so that they can take full advantage of semantic and depth information. The structure of FSM is presented in Fig.4(b). The depth features $\mathbf{fd}_t$ and segmentation features $\mathbf{fs}_t$ are first concatenated, then fed into an architecture resembling encoder-decoder. We utilize $C(\cdot)$ to represent the series of convolutions in the aforementioned process. It can be noticed that we use depthwise separable convolution to reduce computational cost. Eventually the commonly shared features are allocated adaptively into two branches via 1×1 convolutions $C_{\mathbf{fd}}^{1 \times 1}(\cdot)$, $C_{\mathbf{fs}}^{1 \times 1}(\cdot)$. Followed by residual connections, the features $\mathbf{d}_{t+1}$ and $\mathbf{s}_{t+1}$ can be obtained by:

$$\begin{aligned} \mathbf{fd}_{t+1} &= \mathbf{fd}_t + C_{\mathbf{fd}}^{1 \times 1}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))) \\ \mathbf{fs}_{t+1} &= \mathbf{fs}_t + C_{\mathbf{fs}}^{1 \times 1}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))). \end{aligned} \tag{5}$$

We also compare our structure with lateral sharing unit (LSU) proposed in [16][20], which is shown in Fig.4(a). It can be seen that the task-shared features $\mathbf{fs}_{t+1}$ and $\mathbf{fd}_{t+1}$ are obtained with summation of three features, which can be formulated by:

$$\begin{aligned} \mathbf{fd}_{t+1} &= \mathbf{fd}_t + C_{\mathbf{fd1}}^{1 \times 1}(\mathbf{fd}_t) + C_{\mathbf{fs2}}^{1 \times 1}(\mathbf{fs}_t) \\ \mathbf{fs}_{t+1} &= \mathbf{fs}_t + C_{\mathbf{fd2}}^{1 \times 1}(\mathbf{fd}_t) + C_{\mathbf{fs1}}^{1 \times 1}(\mathbf{fs}_t), \end{aligned} \tag{6}$$

Although their method, to some extend, realizes the interaction of different features, providing information for later predictions, we argue that the element-wise summation can only obtain local information, making limited use of the fused features. For example, the depth features located at (i, j) can only sum with the corresponding semantic features. Contrast to LSU, our method implements sampling towards the interacted features, which encapsulates large area of features and augments the presentation ability. Therefore, each task-specific feature acquires more useful information. We insert FSM before each stage of upsampling, benefiting depth split and segmentation split to exploit multi-level fused information.

D. Training Loss Function

Besides the previously mentioned attention loss, our loss function includes three other parts: consistency loss, depth loss and segmentation loss.

Consistency Loss: Inspired by [13], we design a consistency loss to make semantic and depth branches guide each other mutually. Specifically, the features that are distinct or
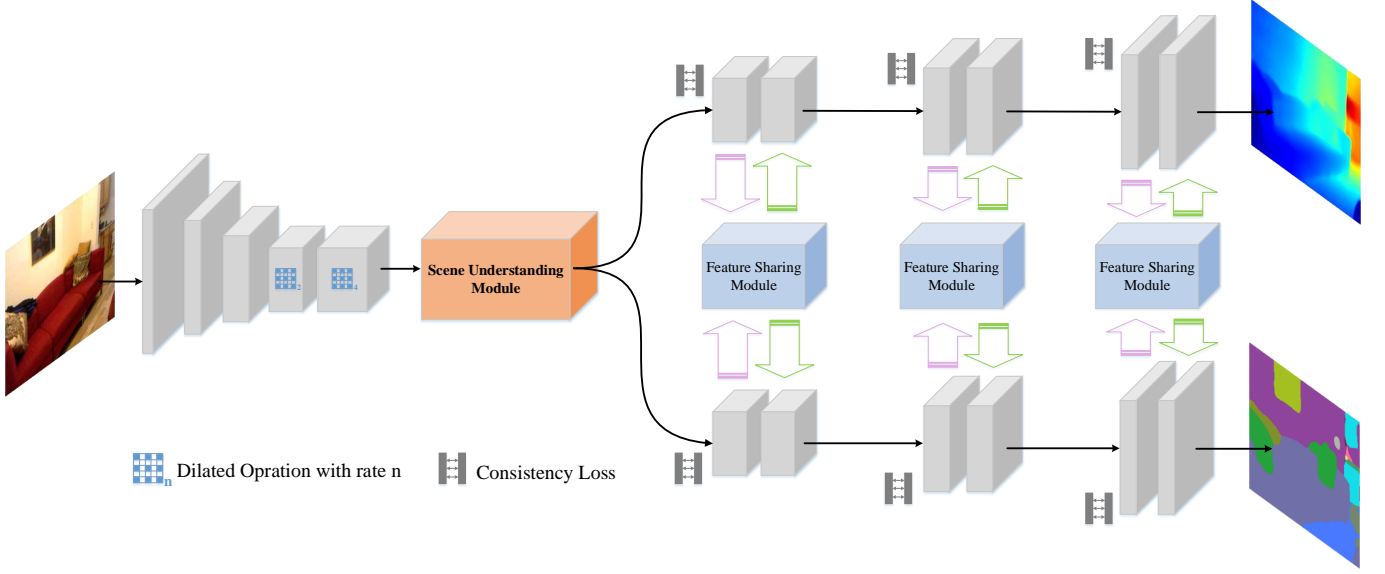
Fig. 1. The overview of our CI-Net for joint depth estimation and semantic segmentation. We adopt dilated operation in the backbone to mitigate the harm of over-downsampling. At the end of the encoder, the SUM is designed to aggregate the contextual information. Then the network breaks into depth and segmentation split. To exploit the task-specific features, the FSM is proposed. Finally, a consistency loss is formulated to make the depth and segmentation features mutually guided.
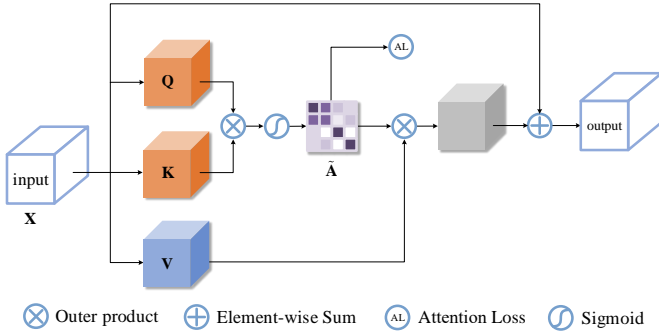

Fig. 2. The structure of SUM. The generated attention map captures context prior of inter-class and intra-class so that both of segmentation features and depth features understand the scene better.
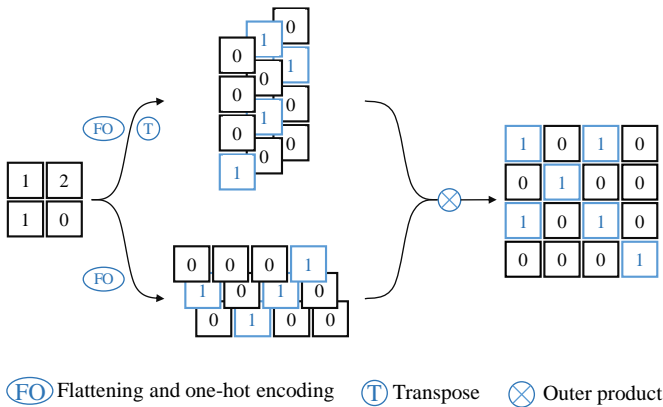

Fig. 3. The process of generating the ideal attention map. First we implement one-hot encoding for the semantic label and then flatten it into a $HW \times 1 \times C$ matrix $\mathbf{M}$ where C denotes the number of categories. After the operation of outer product, the ideal attention map $\mathbf{A}$ is constructed with size of $HW \times HW$. It can be noticed that the pixels of same classes in semantic label have the same value 1 in the attention map.

similar in semantic feature map should keep the same nature as in depth representations. For example, the semantic features of the sky and tree are extremely different since they belong to different classes, while the corresponding two depth features are also discrepant because the distances of them have large gap. Therefore, we employ this character to supervise the consistency loss of task-specific features, the form of which is defined as

$$\mathcal{L}_{con} = \sum_l \sum_{i,j} \psi(s_{i,j}, s_{i,j}(l))|D(\mathbf{fd}_{i,j}, l) - D(\mathbf{fs}_{i,j}, l)|$$

$$D(\mathbf{fd}_{i,j}, l) = \exp[-\frac{1}{2}(\mathbf{fd}_{i,j} - \mathbf{fd}_{i,j}(l))^T \mathbf{\Sigma}_{\mathbf{fd}}^{-1}(\mathbf{fd}_{i,j} - \mathbf{fd}_{i,j}(l))].$$
(7)

where s_{ij} denotes the semantic label of ground truth. $\mathbf{fd}_{i,j}(l)$ is the depth feature which has an offset of l along x or y direction and $\mathbf{fs}_{i,j}(l)$ is the semantic feature. We use the exponential form of Mahalanobis Distance to measure the discrepancies between features where the covariance matrix $\mathbf{\Sigma}_{\mathbf{fd}}$ is set as a diagonal matrix $\sigma^2 \mathbf{I}_C$. Here σ is a learned parameter from each feature map. Considering that the depth features at the inner edges vary widely while the semantic representations are similar, we weight L_{con} by the function $\psi(\cdot)$ which returns 1 when the corresponding labels are different and 0 otherwise. Since it is not realistic to take account of all the feature relationships, we select l from the set $\{1, 2\}$ so that each feature would be compared for 8 times, which has adequately good performance in our experiments.

Depth Loss: The depth loss is composed of three items $\mathcal{L}_{berhu}$, $\mathcal{L}_{pair}$ and $\mathcal{L}_{norm}$. The $\mathcal{L}_{berhu}$ represents the BerHu Loss providing a good balance of the L1 norm and L2 norm, which is effective in the occasion errors following a heavy-
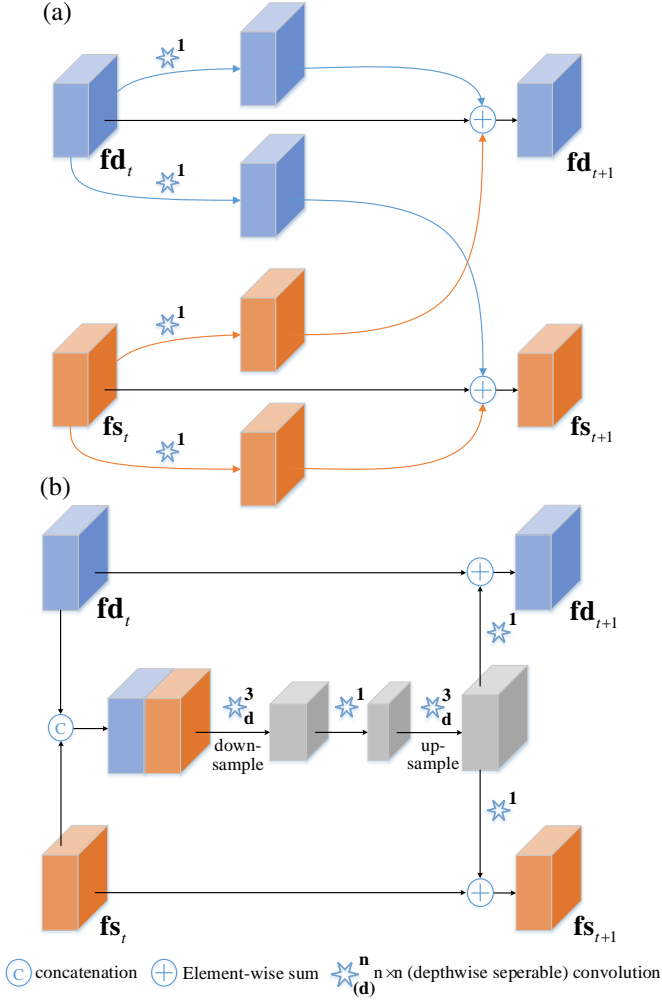
Fig. 4. (a) The architecture of LSU; (b) The architecture of our proposed FSM [16][20]. our module can capture extensive interacted context, while the comparative one only captures local interaction.

tailed distribution [25]. The $\mathcal{L}_{berhu}$ is defined by

$$\mathcal{L}_{berhu} = \sum_{i,j} \begin{cases} |d_{i,j} - \tilde{d}_{i,j}| & \text{if} |d_{i,j} - \tilde{d}_{i,j}| \leq c, \\ \frac{(d_{i,j} - \tilde{d}_{i,j})^2 + c^2}{2c} & \text{if} |d_{i,j} - \tilde{d}_{i,j}| > c \end{cases}, \quad (8)$$

where $d_{i,j}$ and $\tilde{d}_{i,j}$ are respectively the true and estimated depth values. c is a threshold and we set it to be $c = \frac{1}{5} max_k(|d_k - \tilde{d}_k|)$, that is 0.2 times of the max error in a batch.

We also introduce the loss term $\mathcal{L}_{pair}$ to ensure the smoothness in the homogeneous regions and the relative distance of different areas. The formulation of $\mathcal{L}_{pair}$ is

$$\mathcal{L}_{pair} = \sum_{p,q \in \Lambda, p \neq q} ||d_p - d_q| - |\tilde{d}_p - \tilde{d}_q||, \quad (9)$$

in which $\Lambda = \{(i_1, j_1), (i_2, j_2), \ldots, (i_n, j_n)\}$ denotes the set of pixel indices which are selected randomly. We argue that $\mathcal{L}_{pair}$ not only keeps the advantages of the gradient loss [35] which penalizes the adjacent pixels of smooth areas and discontinued borders, but also provides similarity of pixels that are far apart, guaranteeing relative distances of different objects.

Another loss term is $\mathcal{L}_{norm}$, which is employed to emphasize small-size structures and high-frequency details:

$$\mathcal{L}_{norm} = \sum_{i,j} (1 - \frac{\mathbf{n}_{i,j} \cdot \tilde{\mathbf{n}}_{i,j}}{|\mathbf{n}_{i,j}| \cdot |\tilde{\mathbf{n}}_{i,j}|}), \quad (10)$$

where $\mathbf{n}_{i,j}$ is the surface normal calculated by $\mathbf{n}_{i,j} = (-\nabla_x d_{i,j}, -\nabla_y d_{i,j}, 1)^T$, in which ∇_x and ∇_y represents the gradient values along the x-axis and y-axis separately. The depth loss is then calculated by the weighted summation of these three terms:

$$\mathcal{L}_{depth} = \mathcal{L}_{berhu} + \lambda \mathcal{L}_{pair} + \mu \mathcal{L}_{norm}, \quad (11)$$

where λ, μ are weights to balance the depth loss terms.

Segmentation Loss: To ensure the accuracy of semantic segmentation and avoid unfavorable depth estimation along object boundaries, we introduce the weighted cross-entropy loss as segmentation loss $\mathcal{L}_{seg}$. Then our total loss is

$$\mathcal{L} = \mathcal{L}_{att} + \alpha \mathcal{L}_{depth} + \beta \mathcal{L}_{con} + \gamma \mathcal{L}_{seg}, \quad (12)$$

where α, β, γ denotes the weight coefficients for each loss.

IV. EXPERIMENTS

In this section, we first introduce the training datasets and evaluation metrics. Then we illustrate the implementation details of training our model. Next, we investigate the effectiveness of our proposed network and compare it with state-of-the-art methods. Ablation study is also performed to show the benefits of our proposed methods.

A. Dataset and Metrics

Dataset: We use NYU-Depth-v2 [40] and SUN-RGBD [41] datasets to evaluate our presented model. NYU-Depth-v2 includes around 120K RGB-D images of 464 indoor scenes where only 1449 images from them have semantic labels. We follow the methods adopted in [20] [31] to use the standard 795 training pairs and 654 testing pairs. SUN-RGBD is another dataset obtained from the indoor scenes containing about 10K images(5285 images for training and 5050 images for testing). Since the dataset has both semantic labels and depth labels, the whole training set is utilized to train our model and the test set to evaluate our semantic predictions. Both of these two datasets are employed for comparing with the state-of-the-art methods.

Metrics: Like the previous works [4], we assess our predicted depth maps using the following metrics:

Accuracy with threshold (δ^p): % of $d_{i,j}$ s.t. $max(\frac{\tilde{d}_{i,j}}{d_{i,j}}, \frac{d_{i,j}}{\tilde{d}_{i,j}}) = \delta^p < 1.25^p (p = 1, 2, 3)$

RMSE (rms): $\sqrt{\frac{1}{N} \sum_{i,j} (d_{i,j} - \tilde{d}_{i,j})^2}$

RMSE in log space (rms_log): $\sqrt{\frac{1}{N} \sum_{i,j} (\ln d_{i,j} - \ln \tilde{d}_{i,j})^2}$

Mean absolute relative error (abs_rel): $\frac{1}{N} \sum_{i,j} \frac{|d_{i,j} - \tilde{d}_{i,j}|}{d_{i,j}}$

Mean relative square error (sq_rel): $\frac{1}{N} \sum_{i,j} \frac{(d_{i,j} - \tilde{d}_{i,j})^2}{d_{i,j}^2}$

The signal N represents the number of valid pixels.

To evaluate the predictions of semantic segmentation, we refer to the recent work [19][42] and introduce pixel accuracy (pAcc) and mean intersection over union (mIoU) as metrics.

TABLE I

ABLATION STUDY OF OUR METHOD ON THE NYU-DEPTH-V2 DATASET. THE BASELINE MODEL IS OUR MODEL REMOVED FSM AND SUM. BOTH DEPTH ESTIMATION AND SEMANTIC SEGMENTATION RESULTS ARE PRESENTED.

| Backbone	Improvements			Depth					Segmentation	
	SUM	FSM	$\mathcal{L}_{csc}$	higher is better			lower is better		higher is better	
				δ_1	δ_2	δ_3	rms / rms_log	abs_rel / sq_rel	pAcc	mIoU
Dilated ResNet-101				0.769	0.942	0.982	0.563 / 0.208	0.159 / 0.121	67.1	37.2
	✓			0.792	0.953	0.987	0.518 / 0.192	0.142 / 0.115	70.9	41.2
		✓		0.785	0.950	0.988	0.538 / 0.201	0.152 / 0.118	69.5	39.0
	✓	✓		0.809	0.957	0.990	0.512 / 0.187	0.140 / 0.112	72.2	42.5
	✓	✓	✓	0.812	0.957	0.990	0.504 / 0.181	0.129 / 0.112	72.7	42.6
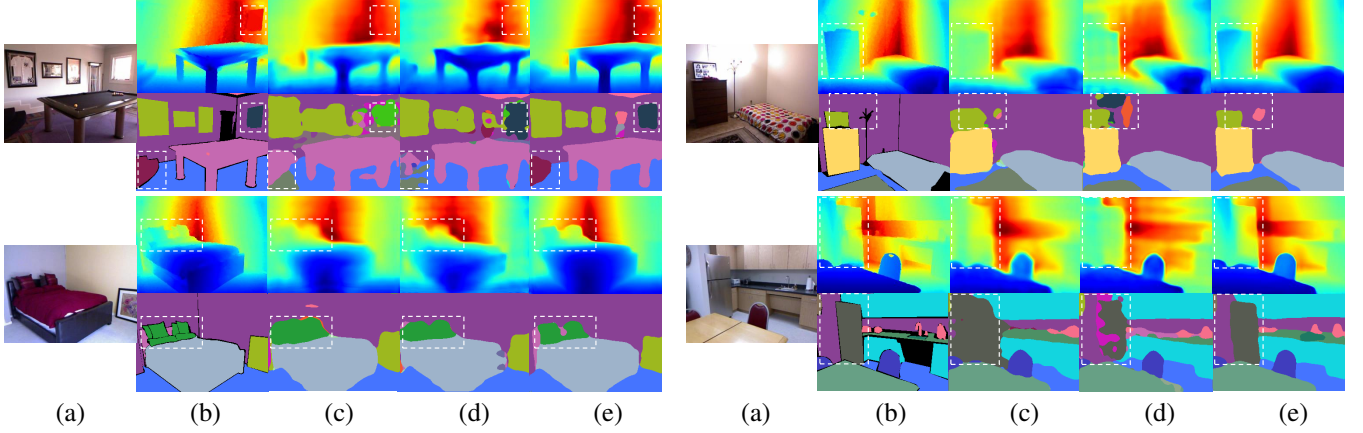
Fig. 5. Ablative visual comparisons. (a) input image; (b) ground truth; (c) results of baseline; (d) results of baseline with FSM; (e) results of our method.

B. Implementation Details

We implement the model using the open source machine learning framework Pytorch on a single Nvidia GTX1080Ti GPU. As for the encoder of CI-Net, we choose the ResNet-50 and ResNet-101 as the candidates and both of them are pretrained on the ImageNet classification task [43]. The learning rates of the pretrained layers are set to be 10 times smaller than the other layers. To avoid the overfitting problem, we adopt the data augmentation strategies, including random rotation, random scaling, random crop, random horizontal flip and random color jitter. The Optimization Algorithm we used is stochastic gradient descent (SGD) [44] where we set the momentum as 0.9 and the weight decay as $5e^{-4}$. To guarantee computational efficiency and fully optimizing the network, we choose the number of set Λ as 500 to compute the pair loss $\mathcal{L}_{pair}$. The weight coefficients $(\alpha, \beta, \gamma, \lambda, \mu)$ are set to $(1, 5, 0.3, 1, 5)$ respectively. The training process is divided into three stages. At first, we replace the SUM module with the ground truth attention maps and the model is trained using only $\mathcal{L}_{depth}$ and $\mathcal{L}_{seg}$ (epochs and learning rates are $(300, 6e^{-4})$ for NYU-Depth-v2 and $(60, 6e^{-4})$ for SUN-RGBD). During the second stage, the SUM is added to the model and L_{att} participates in the training process (epochs and learning rates are $(200, 2e^{-4})$ for NYU-Depth-v2 and $(40, 3e^{-4})$ for SUN-RGBD). In the last stage, we employ all the loss costs to train the entire model (we use the polynomical decay strategy with decayed power of 0.9, epochs and initial learning rates are $(200, 2e^{-4})$ for NYU-Depth-v2 and $(40, 3e^{-4})$ for SUN-

RGBD).

C. Ablation Study

In this subsection, we conduct numerous ablative experiments to analyze the effectiveness of our settings for the model. The experiments are evaluated on NYU-Depth-v2 dataset. To make an ablation study, we set a baseline network comprised of one encoder followed by two separate decoders each of which corresponds to a single task and consists of three blocks. On the basis of baseline, we also train a network with FSM to evaluate the effectiveness of SUM and FSM. Both networks and our proposed model are equipped with dilated ResNet-101 as encoder.

We first analyze the contribution of scene understanding module. The ablative visual results are shown in (d) and (e) of Fig.5. It can be observed that without context prior, both the depth estimation and segmentation results of baseline with FSM suffer noticeable errors especially in the white dashed line boxes. We also observe that SUM can significantly reduce the adverse effect of uneven illumination. For example, in the second scene there is a lamp shedding intense light which impairs the baseline prediction of the surrounding region. In this case, SUM provides the understanding of scene, which helps accurately predict depth and semantic information. In addition, we visualize the attention maps with and without the supervision of attention loss respectively. Fig.6 shows that with the guidance of attention loss, the model does capture the correlated contextual areas and adapts to different scenes

well. The attention map with supervision could be regarded as a structural extractor since it extracts intact object shapes, revealing the layout of scene. Whereas for the attention maps are not properly guided, the resulting arbitrary concerned region can be harmful.
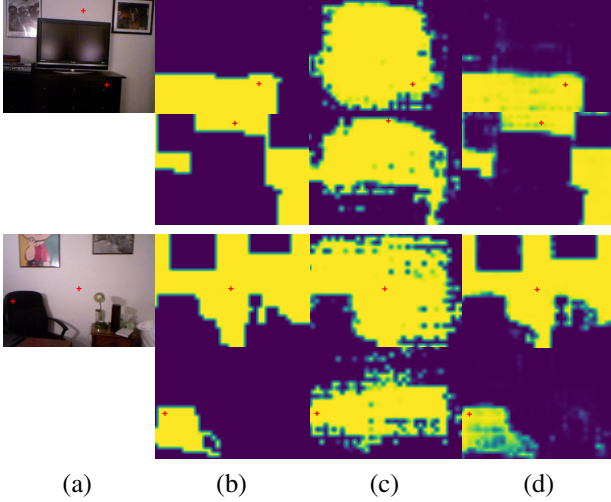

(a)   (b)   (c)   (d)

Fig. 6. Comparisons of the learned attention map. (a) input images; (b) ideal attention maps; (c) and (d) represent the attention maps produced by our model without and with supervision of the $\mathcal{L}_{att}$ respectively. For each scene we show two different attention maps pertain to the locations where the red plus signs mark.

TABLE II
COMPARISONS BETWEEN FSM AND LSU

	δ_1	δ_2	δ_3	rms	abs_rel	pAcc	mIoU
baseline	0.769	0.942	0.982	0.563	0.159	67.1	37.2
baseline + LSU	0.778	0.948	0.986	0.547	0.152	68.3	38.5
baseline + FSM	0.785	0.950	0.988	0.538	0.152	69.5	39.0

Next, it is obvious that the feature sharing strategy improves the performance of both semantic segmentation and depth estimation. For the example in the forth scene of Fig.5, the baseline fails to predict the wall behind fridge while FSM help perceive the existence of these two objects and make boundaries in the depth map clearer. These improvements are facilitated by deeply fusing the task specific features, providing more robust information for prediction. In addition, we replace the FSM with LSU [16][20] whose results can be seen in TABLE II. It is noticed that Although the LSU does improve the performance, the improvements are not as obvious as those by adding FSM, which proves the effectiveness of sampling operations.

Quantitative comparisons are also performed, which can be seen in TABLE I. It can be obviously found that both the introductions of SUM and FSM improve the performance significantly, which verifies the effectiveness of these two blocks. For the case with scene understanding module, i.e., baseline with SUM, the original baseline network could obtain a prominent gain on both tasks, especially in rms and mIoU (7.9% reduced and 10.7% increased respectively). This result agrees well with the qualitative data and indicates that

improving the understanding of scene is effective. Moreover, we also exhibits that when the extra supervision of consistency loss is added, the accuracies are slightly enhanced.
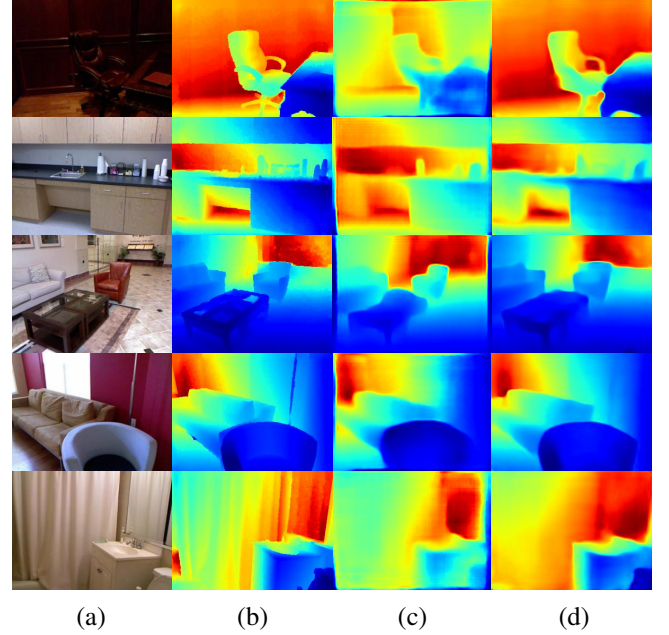

(a)   (b)   (c)   (d)

Fig. 7. Visual comparison with some approaches on NYU-depth-v2 dataset. (a) input images; (b) ground truth; (c) predictions of [25]; (d) predictions of our model.

D. Comparison with State-of-the-arts

In this subsection, we compare the experimental results of our model with other state-of-the-arts according to different tasks.

Depth Estimation: We firstly make comparison of depth values among our presented CI-Net and state-of-the-arts on NYU-Depth-v2 dataset where the number of categories are 40. The results are summarized in TABLE III. The values are used directly from the original papers or referred to the collected data of [39]. For our network, we consider two settings of the backbone: dilated resnet-50 and dilated resnet-101, which are denoted as Res50 and Res101 respectively in the table. As it can be seen in the TABLE III, our model with ResNet-101 achieves best results in all metrics compared to the model using the same scale data. The rms of CI-Net is even better than those methods fed with larger training set. And the vision with dilated ResNet-50 obtains goodish values. It's also worth noting that although the methods in [17][53][54] perform well in part of the evaluating indicators, they all attempted to feed numerous images (120k) to boost the performance and contains large amounts of parameters which hinder the inference speeds of their models. In contrast, ours has least amount of parameters and runs 1.62, 1.45, 4.53 times faster than them respectively.

In addition, Fig.7 displays some visual examples. It can be seen that though the predictions of the method by Laina [25] are smoothed as a whole, they lose some details, bringing in the blurred object boundaries especially in desk, washing machine and sofa. Besides, the precision of depth maps is

TABLE III
COMPARISONS WITH THE DEPTH ESTIMATION METHODS ON NYU-DEPTH-V2 DATASET.

Methods	Data	Params	Memory	FPS	higher is better			lower is better			
					δ_1	δ_2	δ_3	rms	rms_log	abs_rel	sq_rel
Wang et al.[30]	795	/	/	/	0.605	0.890	0.970	0.745	0.262	0.220	0.210
DCNF[45]	795	40.0M	/	/	0.614	0.883	0.971	0.824	/	0.230	/
HCRF[46]	795	/	/	/	0.621	0.886	0.968	0.821	/	0.232	/
Roy et al.[47]	795	/	/	/	/	/	/	0.744	/	0.187	/
PAD-Net [12]	795	/	/	/	0.817	0.954	0.987	0.582	/	0.120	/
FCRN [25]	12k	63.4M	/	/	0.811	0.953	0.988	0.573	0.195	0.127	/
Li et al. [48]	12k	/	/	/	0.820	0.960	0.989	0.545	/	0.139	/
Moukari et al. [49]	16k	/	/	/	0.830	0.966	0.993	0.569	/	0.133	/
GeoNet [50]	16k	67.2M	/	/	0.834	0.960	0.990	0.569	/	0.128	/
Eigen et al.[4]	120k	141.1M	/	/	0.611	0.887	0.971	0.907	0.285	0.158	0.121
Eigen et al.[51]	120k	54.2M	/	/	0.769	0.950	0.988	0.641	0.214	0.158	0.121
CRFs[52]	95k	/	/	/	0.811	0.954	0.987	0.586	/	0.121	/
DORN [17]	120k	110.0M	489.1MB	42	0.828	0.965	0.992	0.509	/	0.115	/
SharpNet [53]	120k	80.4M	336.6MB	47	0.836	0.966	0.993	0.502	/	0.139	/
Hu et al. [54]	120k	157.0M	679.7MB	15	0.866	0.975	0.993	0.530	/	0.115	/
Our(Res50)	795	34.7M	132.4MB	90	0.799	0.952	0.984	0.533	0.201	0.145	0.121
Our(Res101)	795	53.7M	204.8MB	68	0.812	0.957	0.990	0.504	0.181	0.129	0.112

TABLE IV
COMPARISONS WITH THE SEMANTIC SEGMENTATION METHODS ON NYU-DEPTH-V2 DATASET

Methods	Data	pAcc	mIoU
FCN [10]	RGB	60.0	29.2
Context [55]	RGB	70.0	40.6
Eigen et al. [51]	RGB	65.6	34.1
B-SegNet [56]	RGB	68.0	32.4
Deng et al. [42]	RGBD	63.8	31.5
He et al. [57]	RGBD	70.1	40.1
3DGNN [58]	RGBD	/	42.0
D-CNN [8]	RGBD	53.6	41.0
Our(Res50)	RGB	69.1	40.1
Our(Res101)	RGB	72.7	42.6

weak, the depths of some regions deviate the ground truth severely. On the contrary, our results match the structures of the scenes and have sharper object boundaries.

TABLE V
COMPARISONS WITH THE SEMANTIC SEGMENTATION METHODS ON SUN-RGBD DATASET

Methods	Data	pAcc	mIoU
Context [55]	RGB	78.4	42.3
B-SegNet [56]	RGB	71.2	30.7
RefineNet-101 [11]	RGB	80.4	45.7
FuseNet [59]	RGBD	76.3	37.3
D-CNN [8]	RGBD	/	42.0
Our(Res50)	RGB	75.4	42.5
Our(Res101)	RGB	80.7	44.3

Semantic Segmentation: Apart from depth estimation, we show the semantic segmentation results in TABLE IV and TABLE V for NYU-Depth-v2 and SUN-RGBD datasets respectively. As reported in TABLE IV, our model with ResNet-101 performs best among all the listed methods in both pixel-acc and mIoU even though our method only uses RGB images, and the model equipped with ResNet-50 achieves comparable results, which demonstrate that the context prior can also benefit the semantic segmentation and the deep feature interaction does help this task leverage more useful information. As for the results of SUN-RGBD dataset in TABLE V, we can notice that our method still performs competitively. Although RefineNet-101 achieves best performance in mIoU metric, our method is on a par with it in pixel-acc metric. Some selected visual maps of both datasets are also shown in Fig.8, it can be seen that our predictions are in high quality and even give the correct labels of the invalid regions in ground truth.

Fig. 8. Visualized semantic segmentation maps on SUN-RGBD dataset. (a) input images; (b) ground truth; (c) semantic segmentation predictions of our method.

V. CONCLUSION

In this paper, a CI-Net was proposed for joint depth estimation and semantic segmentation. By employing the scene

understanding module, the presented CI-Net is able to capture the contextual information of inter-class and intra-class, which is crucial for the network to understand which useful context can be exploited to make predictions. We also design a feature sharing module to deeply fuse the task-specific features. By putting the shared features through upsampling and down-sampling operations, the network could learn more robust representations. In addition, a consistency loss is proposed to make the task-specific features mutually guided, which keeps consistent relationships with the respective adjacent features. The results of experiments prove the effectiveness of our proposed algorithm. In the future, we plan to explore the contextual information in the attention map and incorporate other pixel-level tasks such as surface normal prediction, edge detection into this work. Also, we are interested in combining the depth prediction with semantic SLAM to obtain more accurate results.

REFERENCES

[1] X. Zhao, J. Liu, X. Wu, W. Chen, F. Guo, and Z. Li, "Probabilistic spatial distribution prior based attentional keypoints matching network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[2] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] F. Tang, Y. Wu, X. Hou, and H. Ling, "3d mapping and 6d pose computation for real time augmented reality on cylindrical objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2887–2899, 2019.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *arXiv preprint arXiv:1406.2283*, 2014.

[5] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3917–3925.

[6] 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, 2015.

[7] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[8] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

[9] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4823–4836, 2019.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[11] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[12] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.

[13] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," *arXiv preprint arXiv:2002.12319*, 2020.

[14] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4106–4115.

[15] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 235–251.

[16] J. Choi, D. Jung, D. Lee, and C. Kim, "Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction," *arXiv preprint arXiv:2010.02893*, 2020.

[17] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[18] Y. Chen, H. Zhao, Z. Hu, and J. Peng, "Attention-based context aggregation network for monocular depth estimation," *International Journal of Machine Learning and Cybernetics*, pp. 1–14, 2021.

[19] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 416–12 425.

[20] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 53–69.

[21] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *European Conference on Computer Vision*. Springer, 2020, pp. 582–600.

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

[23] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[24] A. Saxena, S. H. Chung, A. Y. Ng *et al.*, "Learning depth from single monocular images," in *NIPS*, vol. 18, 2005, pp. 1–8.

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

[26] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5684–5693.

[27] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[28] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 225–234.

[29] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[30] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2800–2809.

[31] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[35] T. Chen, S. An, Y. Zhang, C. Ma, H. Wang, X. Guo, and W. Zheng, "Improving monocular depth estimation by leveraging structural awareness and complementary datasets," *arXiv preprint arXiv:2007.11256*, 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472–480.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[39] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu, and J. Heikkilä, "Guiding monocular depth estimation using depth-attention volume," in *European Conference on Computer Vision*. Springer, 2020, pp. 581–597.

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

[41] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[42] Z. Deng, S. Todorovic, and L. Jan Latecki, "Semantic segmentation of rgbd images with mutex constraints," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1733–1741.

[43] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-d scene structure from a single still image," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[44] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[45] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[46] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1119–1127.

[47] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5506–5514.

[48] B. Li, Y. Dai, and M. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2018.

[49] M. Moukari, S. Picard, L. Simon, and F. Jurie, "Deep multi-scale architectures for monocular depth estimation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2940–2944.

[50] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 283–291.

[51] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[52] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[53] M. Ramamonjisoa and V. Lepetit, "Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[54] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1043–1051.

[55] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3194–3203.

[56] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[57] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, "Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4837–4846.

[58] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5199–5208.

[59] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*. Springer, 2016, pp. 213–228.