Applied Intelligence
https://doi.org/10.1007/s10489-022-03401-x

CI-Net: a joint depth estimation and semantic segmentation network using contextual information

Tianxiao Gao[1] · Wu Wei[1] · Zhongbin Cai[1] · Zhun Fan[2] · Sheng Quan Xie[3] · Xinmei Wang[4] · Qiuda Yu[1]

Accepted: 16 February 2022

Abstract
Monocular depth estimation and semantic segmentation are two fundamental goals of scene understanding. Due to the advantages of task interaction, many works have studied 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, which limits the performance of both tasks. In this paper, we propose a network injected with contextual information (CI-Net) to solve this problem. Specifically, we introduce a self-attention block in the encoder to generate an attention map. With supervision from the ideal attention map created by semantic label, the network is embedded with contextual information so that it could understand the scene better and utilize correlated features to make accurate prediction. Besides, a feature-sharing module (FSM) is constructed to make the task-specific features deeply fused, and a consistency loss is devised to ensure that the features mutually guided. We extensively evaluate the proposed CI-Net on NYU-Depth-v2, SUN-RGBD, and Cityscapes datasets. The experimental results validate that our proposed CI-Net could effectively improve the accuracy of semantic segmentation and depth estimation.

Keywords Depth estimation · Semantic segmentation · Attention mechanism · Task interaction

1 Introduction

Scene understanding is an important yet challenging task in computer vision and has contributed to visual simultaneous localization and mapping(vSLAM) system [1], robot navigation [2], autonomous driving [3], and other applications. Two fundamental goals of scene understanding are monocular depth estimation [4, 5] and semantic segmentation [6–8], which have been extensively researched by utilizing deep learning. Recently, some works [4, 9, 10] have observed the interactions between these two tasks and utilized their common characteristics to improve each other, achieving substantial performance increases. However, most of these studies used a deep structure as encoder such as ResNet-101 [11], ResNet-50 [10] and SE-ResNet, introducing a large number of downsampling and stride operations, which had a negative influence [12] on depth estimation and semantic segmentation, where fine-grained information is crucial. Despite some works also adopt strategies such as skip-connection [13], up-projection [11] and multi-scale training loss [10] to mitigate this problem, these schemes have significant demands on computation and memory.

Another shortcoming is that current works about joint learning did not 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 a limited contribution to scene understanding of the network. In [14], Chen et al. pointed out that how to obtain the correlation of inter and intra-objects is crucial for depth estimation. Yu et al. [15] also argued that such context makes feature representation more robust for semantic segmentation. Therefore, could we excavate the information of labels more deeply to assist the network modeling of such correlation? Moreover, most approaches achieve task interaction through adding pixel-wise features [13], simply sharing encoder commonly [16] or sharing parameters of convolutional layers [10]. Although these methods leveraged the correlation between tasks, their approaches for fusing features were rough. For example, in [13], the authors fused the features via direct addition and then added them to task-specific features. This simple structure may make it more difficult for

✉ Wu Wei
eeweiwu@126.com

Extended author information available on the last page of the article.

Published online: 09 April 2022

the network to learn more useful representations of the shared features.

To overcome these problems, this paper presents a network injected with contextual information (CI-Net). We adopt a 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 belonging to the same classes and the differences of those pertaining to different classes. Specifically, we introduce a self-attention block [6] to generate an attention map and exploit the semantic labels to create the ideal attention map indicating whether a pair of pixels belongs to the same classes or not. The attention training loss injects the contextual prior into the network, ensuring the structure to use correlated features for more accurate prediction. To make these two tasks deeply interacting, we present two approaches. The first one is that we design a feature-sharing module (FSM). Rather than simply adding the task-specific features, we concatenate and put them through a series of downsampling and upsampling operations, enabling more useful representations to be obtained. We also devise a consistency loss between the depth and semantic features, forcing them to maintain the intrinsic consistency of first-order relationships.

To summarize, the contributions of this paper involve three aspects:

- We propose a dilated network embedded with a SUM with contextual supervision to inject contextual prior about the correlation of inter and intra-classes, predicting both the depth and semantic segmentation maps.
- We construct an FSM to deeply fuse the task-specific features and put forward a consistency loss to maintain the respective features consistent in the relationship with adjacent features.
- Extensive experiments are performed to demonstrate the effectiveness of our methods. Furthermore, the proposed model achieves competitive results against other approaches of depth estimation and semantic segmentation on NYU-Depth-v2, SUN-RGBD, and Cityscapes datasets.

2 Related works

2.1 Monocular depth estimation

With the appearance of convolutional neural networks (CNNs), monocular depth estimation has been thoroughly studied in recent years. Laina et al. [17] proposed a fully convolutional residual architecture and four up-sampling models to restore the resolution of depth maps. Since then, this technique has been developed to a significant degree. Qi et al. proposed GeoNet, which performed joint depth and surface normal map prediction. This work utilized the geometric constraint between normal and depth to train the network, achieving excellent performance on both surface normal and depth estimation. To reduce the information loss induced by excessive pooling, Fu et al. [12] employed atrous spatial pyramid pooling (ASPP) and presented an ordinal loss to model the depth prediction as an ordinal regression problem. Inspired by [12], Chen et al. [14] proposed soft ordinal inference to exploit the predicted probabilities of the whole depth intervals and replaced ASPP with a self-attention module to capture the global context. Recently, Yin et al. [18] projected the depth map to obtain a 3D point cloud, exploiting the loss between virtual normal and ground truth to train the model, an approach that significantly improved the accuracy. To obtain high-quality depth estimation, Ye et al. presented DP-Net, which fuses multi-level features of a designed dual-branch depth estimation model. Some other works [19, 20] also employed the geometric constraints of the consective image sequence to complete unsupervised depth prediction.

2.2 RGB-D semantic segmentation

The outstanding work proposed by Long et al. [8], fully convolutional network (FCN) achieved great improvement of semantic segmentation. Since then, many scholars [15, 21] have researched on this scene-understanding task using only RGB images. After the RGB-D dataset was released, some approaches [22–24] discovered that fusing depth images could significantly improve the segmentation results. Recently, Hu et al. [25] proposed the attention complementary network which fuses weighted depth and semantic features in the encoder. The fusion implementation enabled ACNet to exploit more high-quality features. Hung et al. [26] designed LDFNet, which is also a fusion-based network. Its novelty of incorporating luminance, depth, and color information produced substantial success in semantic segmentation. To reduce the inference time, Chen et al. [27] proposed spatial information guided convolution (S-Conv) which extracts geometric information as convolutional weights and infers the sampling offset according to the 3D spatial information. Different from these algorithms that aimed to improve semantic segmentation with the facilitation of RGB-D images, we design a joint-task learning network to boost both depth estimation and semantic segmentation with only RGB images as input through the deep interaction of each task.

2.3 Joint semantic segmentation and depth estimation

Due to the common nature of pixel-level prediction among different tasks, some works have paid attention to studying joint learning. In [28], a network named C-DCNN was proposed by Liu et al. which added a designed point-wise bilinear layer to fuse the semantic and depth information to produce higher-order features. Jiao et al. [13] 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. [29], using four intermediate auxiliary tasks and providing abundant information for prediction. In recent research, the SOSD-Net [30] made full use of the geometric relations between depth estimation and semantic segmentation for training. Although these works achieved outstanding performance, they did not exploit the feature that semantic labels could help the network capture prior contextual knowledge of the scene to improve the accuracy of prediction.

2.4 Attention mechanism

The attention mechanism has been widely used in CNNs because it allows the network to ignore parts of the input and focus on others. A profusion of attention methods have been designed recently mainly including channel attention [31], spatial attention [32], and self-attention [6]. Inspired by these methods, some works that incorporated the attention mechanism into depth prediction have emerged. Chen et al. [33] interpolated a channel attention block into the encoder and spatial block into the decoder to avoid losing structural information. Xu et al. [4] presented a fused CRF model guided by multi-scale attention. In 2019, Zhang et al. [10] proposed to fuse the attention maps of three different tasks, and then task-specific propagation was performed to spread the attention map to different tasks, effectively improving the accuracy.

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 FSMs 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 be consistent and we also construct a novel consistency loss.

3 Methods

This section illustrates our proposed method for joint semantic segmentation and depth estimation from a 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.

3.1 Network architecture

The proposed contextual information Network (CI-Net) uses the encoder-decoder scheme as shown in Fig. 1. For the encoder, we choose the ResNet [34] for its identity-mapping tackling of the vanishing gradient problem in deeper networks. Another benefit of ResNet is its large receptive field [17], which is a crucial factor for depth estimation and semantic segmentation. However, rather than deploying ResNet as the encoder directly, as in [11], we also adopt a dilation strategy [35] to mitigate the negative effect of overdownsampling in ResNet, an issue that may hinder the predictions of fine-grained depth and semantic maps. With the last two 2 and 4-dilated residual blocks, the original resolution is lowered to $1/8$ instead of $1/32$, reducing the detailed information loss.

In the decoder, a SUM with a supervised attention block is designed to fully exploit the semantic labels to obtain a context prior of inter and intra-classes, which benefits the model to understand a scene better for later prediction. The network then breaks apart into two branches for estimating depth and segmenting semantics. During this stage, we present an FSM to share feature representations so that these two splits could fully exploit different levels of features. Furthermore, a consistency loss is formulated to keep the task-specific features consistent. More details about these methods are described in the following subsections.

3.2 Scene-understanding module

Our motivation mainly comes from two areas: i) Pixels of the same objects tend to have continuous or similar depth values, whereas the depths of different objects have large discrepancies. ii) Under the background of joint-task learning of semantic segmentation and depth estimation, semantic labels contain information of each class so that it is easy to know whether pixels belong to the same classes or not. Thus, our goal is to find an effective way to ensure the network has prior knowledge of the categorical relationship. To achieve this aim, we utilize the semantic labels 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 prior knowledge of the scene, the profitable information for prediction could be searched in a limited, related space instead of the entire 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

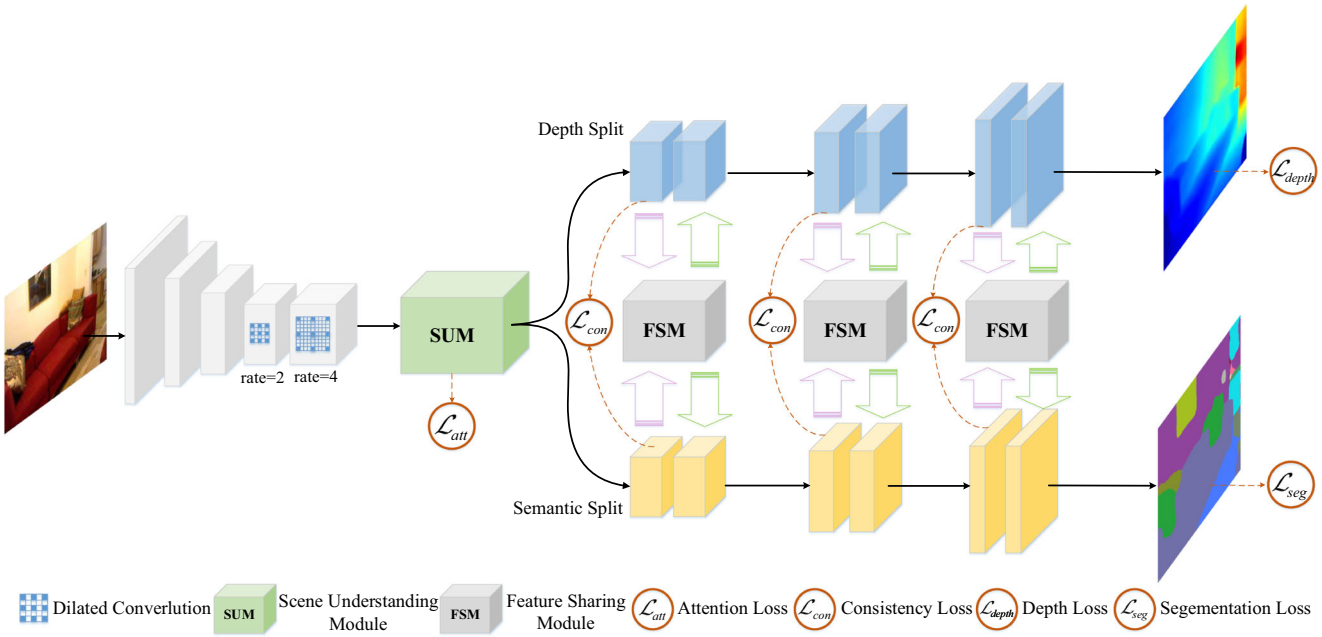

Fig. 1 The overview of our CI-Net for joint depth estimation and semantic segmentation. We adopt dilated operation in the backbone to mitigate the harm of over-downsampling. At the end of the encoder, the SUM is designed to aggregate the contextual information. Then the network breaks into depth and segmentation split. To deeply fuse the task-specific features, the FSM is proposed. Finally, a consistency loss is formulated to make the depth and segmentation features mutually guided

by the context prior for the gap between these two objects. On the other hand, the semantic one also benefits because it makes better judgments from the information of inter and intra-classes. We encapsulate this process of obtaining contextual information in the SUM, which is illustrated in the following content.

The architecture of the SUM is presented in Fig. 2. Similar to non-local block [6], 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 $Sigmoid(\bullet)$ 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 adding 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 [15] to generate an ideal attention map. As 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

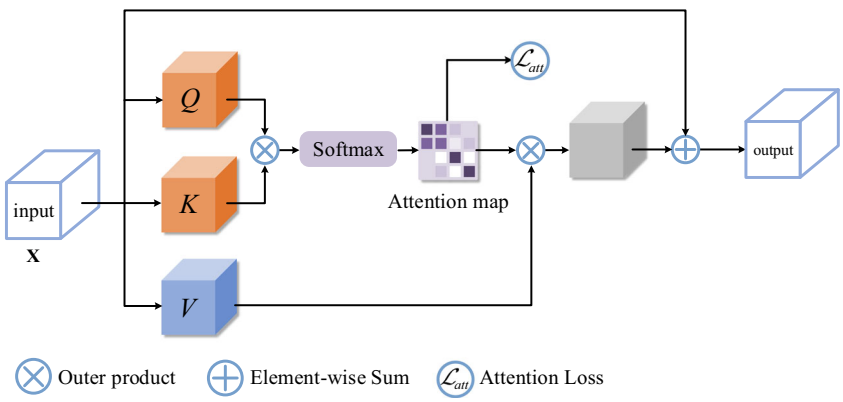

Fig. 2 The structure of SUM. The generated attention map captures context prior of inter and intra-classes so that the network understands the scene better

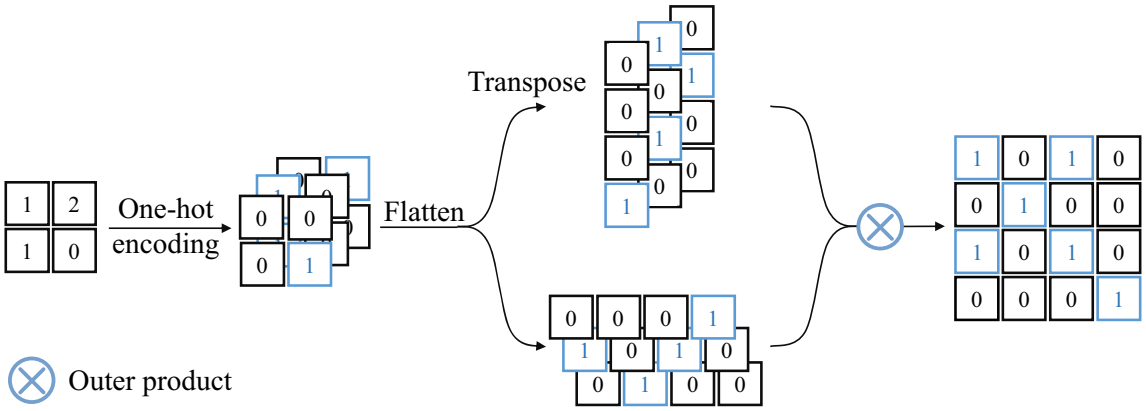

Fig. 3 Process of generating the ideal attention map. First, we implement one-hot encoding for the semantic label and then flatten it into a $HW \times 1 \times C$ matrix $\mathbf{M}$, where C denotes the number of categories. After the operation of outer product, the ideal attention map $\mathbf{A}$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

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 a 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 is 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 and inter-classes. Furthermore, 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 location (i, j) of the predicted attention map.

It is worth noting that we utilize semantic labels rather than depth to inject the context prior. One reason for this approach 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 [14] or planar structures [36], their methods are limited to only depth estimation task. However, for joint-task learning, the correlation of different objects is harmful to semantic segmentation.

3.3 Feature-sharing module

In the decoder, the network splits into depth and semantic branches. We design a 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 FSM structure is presented in Fig. 4b. The depth features $\mathbf{fd}_t$ and segmentation features $\mathbf{fs}_t$ are first concatenated and then fed into an architecture resembling an encoder-decoder. We utilize $C(\bullet)$ to represent the series of convolutions in the aforementioned process. It can be noticed that we use depthwise separable convolution to reduce the computational cost. Eventually, the commonly shared features are allocated adaptively into two branches via 1×1 convolutions $C^{1\times1}_{\mathbf{fd}}(\bullet)$, $C^{1\times1}_{\mathbf{fs}}(\bullet)$. Followed by residual connections, the features $\mathbf{d}_{t+1}$ and $\mathbf{s}_{t+1}$ can be obtained by:

$$\begin{aligned} \mathbf{fd}_{t+1} &= \mathbf{fd}_t + C^{1\times1}_{\mathbf{fd}}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))), \\ \mathbf{fs}_{t+1} &= \mathbf{fs}_t + C^{1\times1}_{\mathbf{fs}}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))). \end{aligned} \tag{5}$$

We also compare our structure with the lateral sharing unit (LSU) proposed in [13], which is shown in Fig. 4a. 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:

$$\begin{aligned} \mathbf{fd}_{t+1} &= \mathbf{fd}_t + C^{1\times1}_{\mathbf{fd1}}(\mathbf{fd}_t) + C^{1\times1}_{\mathbf{fs2}}(\mathbf{fs}_t), \\ \mathbf{fs}_{t+1} &= \mathbf{fs}_t + C^{1\times1}_{\mathbf{fd2}}(\mathbf{fd}_t) + C^{1\times1}_{\mathbf{fs1}}(\mathbf{fs}_t). \end{aligned} \tag{6}$$

Although their method, to some extent, 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. Contrasting with LSU, our method implements sampling towards the interacted features, which encapsulates a large area of features and augments the presentation ability. Therefore, each task-specific feature acquires more useful information. We insert an FSM before each stage of upsampling, benefiting depth split and segmentation split to exploit multi-level fused information.

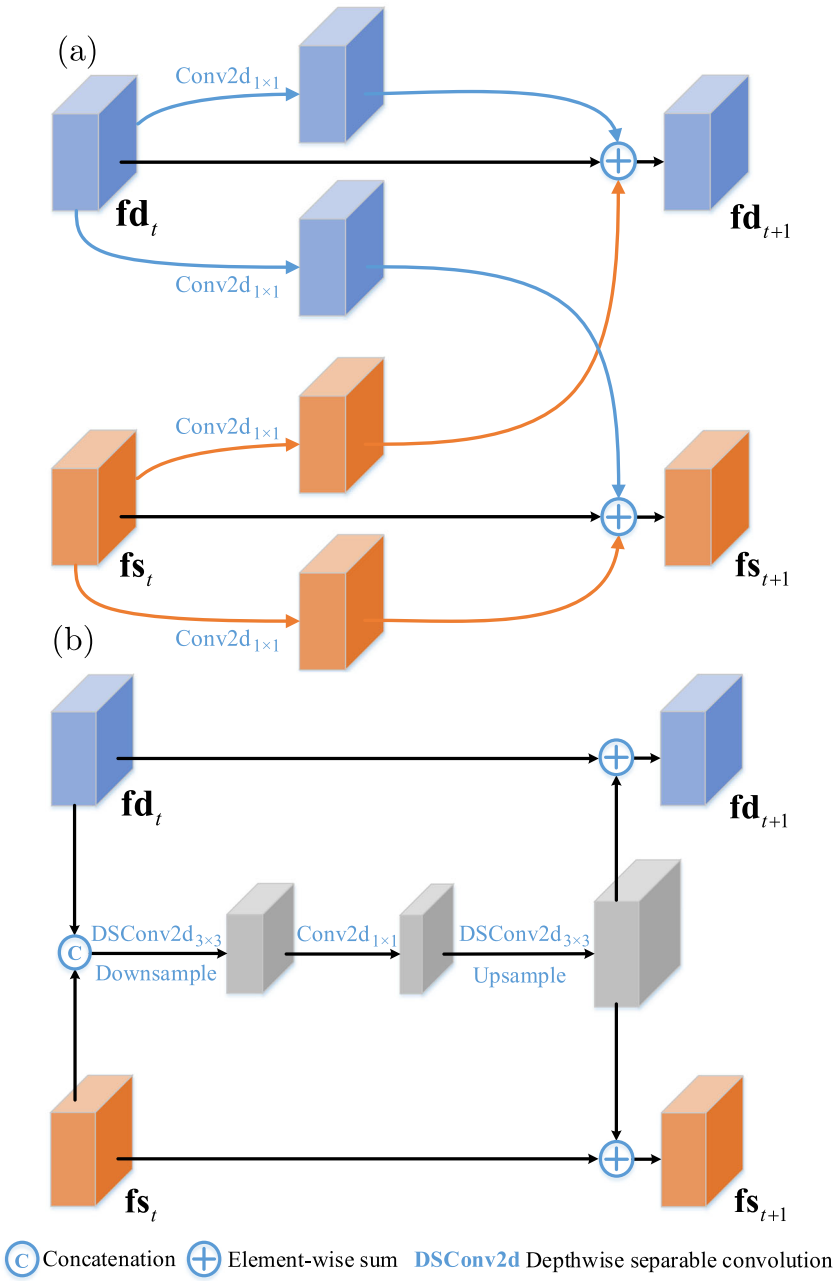

Fig. 4 **a** Architecture of LSU [13]; **b** Architecture of our proposed FSM. Our module can capture extensive interacted context, whereas the comparative one only captures local interactions

3.4 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 [9], we design a consistency loss to make semantic and depth branches guide each other mutually. Specifically, the features that are distinct or similar in semantic feature map should maintain the same characteristics as in the depth representations. For example, the semantic features of the sky and tree are extremely different because they belong to different classes, whereas the corresponding two depth features are also discrepant because the distances between the sky and tree are large. Therefore, we employ this characteristic to supervise the consistency loss of task-specific features, the form of which is defined as

$$\begin{aligned}\mathcal{L}_{con} &= \sum_{l}\sum_{i,j}\psi(s_{i,j}, s_{i,j}(l))|D(\mathbf{fd}_{i,j}, l) - D(\mathbf{fs}_{i,j}, l)|,\\ D(\mathbf{fd}_{i,j}, l) &= \exp[-\tfrac{1}{2}(\mathbf{fd}_{i,j} - \mathbf{fd}_{i,j}(l))^T \Sigma_{\mathbf{fd}}^{-1}\\ &\quad (\mathbf{fd}_{i,j} - \mathbf{fd}_{i,j}(l))],\end{aligned} \tag{7}$$

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 the 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 $\Sigma_{\mathbf{fd}}$ 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, whereas the semantic representations are similar, we weight L_{con} by the function $\psi(\bullet)$ which returns 1 when the corresponding labels are different and 0 otherwise. Because it is not realistic to consider all the feature relationships, we select l from the set $\{1, 2\}$ so that each feature would be compared 8 times, an approach that has adequately good performance in our experiments.

Depth Loss: The depth loss comprises 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-tailed distribution [17]. The $\mathcal{L}_{berhu}$ is defined by

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

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 smoothness in the homogeneous regions and the relative distance of different areas. The formulation of $\mathcal{L}_{pair}$ is

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

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 maintains the advantages of the gradient loss [33], 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:

$$\mathcal{L}_{norm} = \sum_{i,j} 1 - \frac{\mathbf{n}_{i,j} \bullet \tilde{\mathbf{n}}_{i,j}}{|\mathbf{n}_{i,j}| \bullet |\tilde{\mathbf{n}}_{i,j}|}, \tag{10}$$

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 represent the gradient values along the x and y-axes separately. The depth loss is then calculated by the weighted summation of these three terms:

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

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 $\mathcal{L}_{seg}$:

$$\mathcal{L}_{seg} = -\sum_{i,j} \sum_{c} \omega_c \psi(s_{i,j}, c) log(p(\tilde{s}_{i,j}, c)), \tag{12}$$

where $\omega_c = \frac{N_{total} - N_c}{N_{total}}$ weights the segmentation loss to mitigate the category imbalance problem. $p(\tilde{s}_{i,j}, c)$ is the predicted probability value of class c. Then, our total loss is

$$\mathcal{L} = \mathcal{L}_{att} + \alpha \mathcal{L}_{depth} + \beta \mathcal{L}_{con} + \gamma \mathcal{L}_{seg}, \tag{13}$$

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

4 Experiments

In this section, we first introduce the training datasets and evaluation metrics. We then illustrate the implementation details of training our model. Next, an ablation study is performed to show the benefits of our proposed methods. We also investigate the effectiveness of our proposed network and compare it with other methods.

4.1 Dataset and metrics

Dataset: Because both semantic and depth labels are required to train our proposed network, we use three datasets, NYU-Depth-v2 [37], SUN-RGBD [38], and Cityscapes [39] to evaluate our presented model. The popular NYU-Depth-v2 dataset includes approximately 120K RGB-D images of 464 indoor scenes, 1449 images of which have both semantic and depth annotations. We follow the methods adopted in [13, 30] to use the standard 795 training pairs and 654 testing pairs. SUN-RGBD is another commonly used dataset for which images are captured from the indoor scenes containing about 10K images (5285 images for training and 5050 images for testing). Because the dataset has both semantic and depth labels, the entire training set is utilized to train our model and the test set to evaluate our semantic predictions. The outdoor dataset, Cityscapes, focuses on scene understanding of urban street scenes. Collected from over 50 different cities, this dataset has 2975, 500, and 1525 images for training, validation, and test sets respectively. Inverse depth images, instance and semantic segmentation labels are provided for training and testing the network. We transform the inverse depth images into depth maps using the provided camera intrinsic. All of these datasets are employed for ablation and comparison experiments to demonstrate the effectiveness of our proposed method.

Metrics: Similar to the previous works [40], 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 works [15] and introduce pixel accuracy (pAcc) and mean intersection over union (mIoU) as metrics.

4.2 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-101 as the candidates, and both of them are pretrained on the ImageNet classification task. 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 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) [41], 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, $(60, 6e^{-4})$ for SUN-RGBD and $(50, 4e^{-4})$ for Cityscapes). During the second stage, the SUM is added to the model, and $\mathcal{L}_{att}$ participates in the training process (epochs and learning rates are $(200, 2e^{-4})$ for NYU-Depth-v2, $(40, 3e^{-4})$ for SUN-RGBD and $(30, 2e^{-4})$ for Cityscapes). 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, $(40, 3e^{-4})$ for SUN-RGBD and $(30, 2e^{-4})$ for Cityscapes).

4.3 Ablation study

In this subsection, we conduct exhaustive ablative experiments to analyze the effectiveness of our settings for the model. The experiments are extensively evaluated on NYU-Depth-v2, SUN-RGBD, and CityScapes datasets. To show the improvement of SUM and FSM, we set a baseline model comprising one encoder followed by two task-specific decoders, respectively, for depth estimation and semantic segmentation, each of which contains three upsampling blocks (i.e., baseline network: the version of CI-Net removed SUM and FSM). The training loss is a linear combination of task-specific loss (i.e., $\mathcal{L}_{depth}$ and $\mathcal{L}_{seg}$). According to our proposed methods, we trained improved versions of the baseline network, including: **i)** baseline with SUM; **ii)** baseline with FSM; **iii)** baseline with SUM and FSM ; **iv)** CI-Net (introduce $\mathcal{L}_{con}$ to train the version **iii)**). Followed by [14], we use dilated ResNet-101 as the encoder to perform the ablation study.

4.3.1 Scene-understaning module

We first evaluate the contribution of the scene-understanding module (SUM). To better display the improvement, we select some ablative visual results on NYU-Depth-v2, SUN-RGBD, and Cityscapes datasets, which are shown in (d) and (e) of Figs. 5 and 6. It can be observed that without context prior, both depth estimation and segmentation results of the baseline with FSM suffer from noticeable errors particularly in the white dashed line boxes. We also observe that the 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 the scene, which helps accurately predict depth and semantic information. In the outdoor scenes of Cityscapes dataset (Fig. 6), the introduction of the SUM also cleary improves the performance of both semantic segmentation and depth estimation. To take the second scene as an example, when adding the SUM, the predicted depth values of the bus are more accurate and smoothed, whereas without the SUM, the depths of the bus are not sufficiently clear, verifying that the contextual information benefits the network to make high-quality predictions.

Meanwhile, we also perform a quantitative ablative experiment, which can be seen in Table 1. It can be clearly seen that the designed SUM could improve the performance significantly, which verifies the effectiveness of this module. For the NYU-Depth-v2 dataset, the original baseline network could obtain a prominent gain on both tasks, particularly in rms and mIoU (8.0% reduced and 10.7% increased, respectively). The improvement on the SUN-RGBD dataset also agrees well with that on NYU-Depth-v2. It can be observed that the SUM improves the metrics of δ_1, δ_2, δ_3, rms, rms_log, abs_rel, and sq_rel by 0.014, 0.006, 0.005, 0.039, 0.014, 0.015, and 0.006

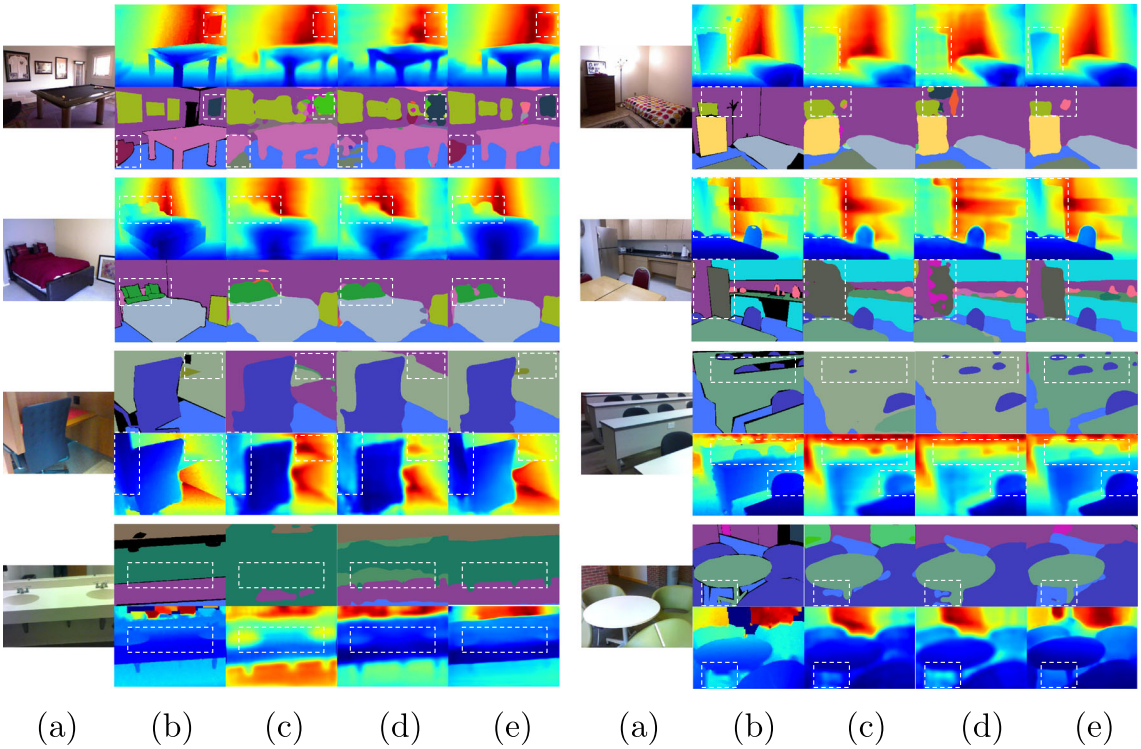

Fig. 5 Ablative visual comparisons on NYU-Depth-v2 and SUN-RGBD datasets. **a** input image; **b** ground truth; **c** results of baseline; **d** results of baseline with FSM; **e** results of our method

respectively. For outdoor scenes of the Cityscapes dataset, the version of baseline with SUM also outperforms the baseline model both on depth estimation (0.781 vs. 0.752 for δ_1, 0.901 vs. 0.882 for δ_2, 0.943 vs. 0.928 for δ_3, 7.021 vs. 7.490 for rms, 0.412 vs. 0.441 for rms_log, 0.238 vs. 0.251 for abs_rel, and 3.966 vs. 4.159 for sq_rel) and semantic segmentation (91.5 vs. 88.0 for pAcc and 67.9 vs. 63.6 for mIoU).

In addition, to see the importance of attention loss towards the SUM, we visualize the attention maps with and without the supervision of attention loss, respectively. Figure 7 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 because it extracts intact object shapes, revealing the layout of a scene. In contrast, as for the attention maps that are not properly guided, the resulting arbitrary concerned region can be harmful.

4.3.2 Feature-sharing module

Next, we verify the effectiveness of the FSM in boosting the performance of depth estimation and semantic segmentation. To take the forth scene of Fig. 5c and d as an example, the baseline fails to predict the wall behind fridge, whereas the FSM helps to 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. The quantitative results are shown in Table 1; it is clear that the version of baseline with FSM achieves better performance than the baseline model. For the NYU-Depth-v2 dataset, compared with the baseline model, baseline with FSM improves all the metrics of depth estimation and semantic segmentation. δ_1, δ_2, δ_3, rms, rms_log, abs_rel, and sq_rel are improved respectively by 0.016, 0.008, 0.006, 0.025, 0.007, 0.007, and 0.003. Both the results of SUN-RGBD and Cityscapes datasets agree well with that on the NYU-Depth-v2 dataset, hence verifying the effectiveness of the FSM.

In addition, to demonstrate the novelty of the FSM, we also compare the performance of the FSM and LSU [13], the results of which can be seen in Table 2. It is noticed that although the LSU does improve the performance, the improvements are not as clear as those by adding the FSM, which illustrates the effectiveness of sampling operations. To make a deep analysis of the difference, we visualize the allocated features $\mathcal{F}_{LSUd}$, $\mathcal{F}_{LSUs}$, $\mathcal{F}_{FSMd}$, and $\mathcal{F}_{FSMs}$(summed long the channel dimension and

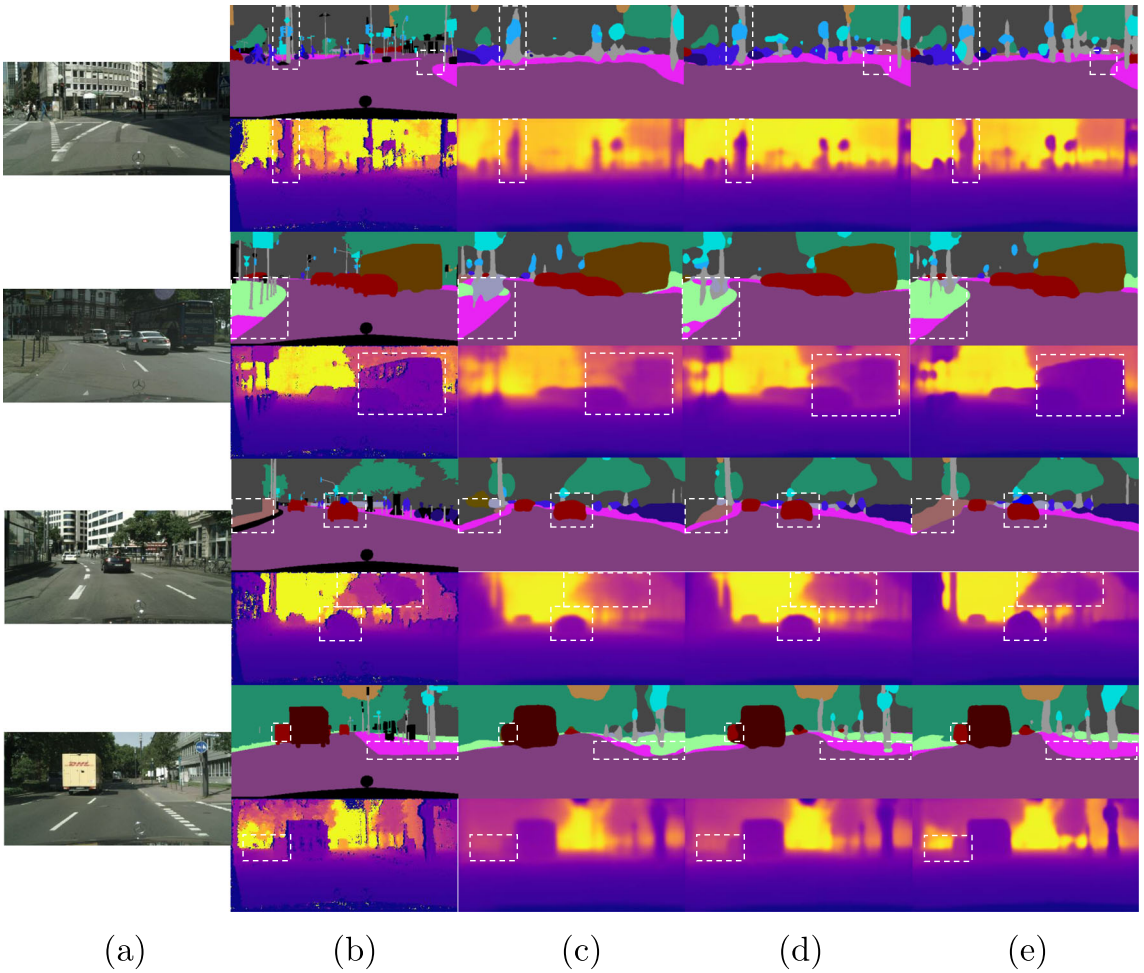

Fig. 6 Ablative visual comparisons on Cityscapes dataset. **a** input image; **b** ground truth; **c** results of baseline; **d** results of baseline with FSM; **e** results of our method

normalized to $0-1$ range) of the LSU and FSM, which are respectively formulated as:

$$\begin{aligned} \mathcal{F}_{LSUd} &= C_{\mathbf{fd1}}^{1\times1}(\mathbf{fd}_t) + C_{\mathbf{fs2}}^{1\times1}(\mathbf{fs}_t) \\ \mathcal{F}_{LSUs} &= C_{\mathbf{fd2}}^{1\times1}(\mathbf{fd}_t) + C_{\mathbf{fs1}}^{1\times1}(\mathbf{fs}_t) \\ \mathcal{F}_{FSMd} &= C_{\mathbf{fd}}^{1\times1}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))) \\ \mathcal{F}_{FSMs} &= C_{\mathbf{fs}}^{1\times1}(C(concat(\mathbf{fd}_t, \mathbf{fs}_t))) \end{aligned} \tag{14}$$

In Fig. 8, we can easily observe that the features learned by the LSU almost pay attention to the entire region, which is not a reasonable approach in depth estimation and semantic segmentation because if all the features of different objects are emphasized, the classification and depth estimation of objects would be confused for using highlighted features from others. In contrast, with larger receptive fields and a deeper structure, our proposed module could learn more useful information, such as objects, which are important in both tasks. In addition, it could be observed that FSM learns clear black boundaries between different objects; when the features are used for lateral convolution, such contours of zero values would inhibit the convoluted operation from using the features of other objects, avoiding the generation of confused features.

4.3.3 CI-Net

When both the SUM and FSM are introduced into the baseline network, the accuracy of both depth estimation and semantic segmentation is drastically improved. For the NYU-Depth-v2 dataset, the most improved metrics are rms and mIoU (9.1% reduced and 14.2% improved, respectively). A similar improvement is also achieved on the SUN-RGBD and Cityscapes datasets, demonstrating that our proposed CI-Net could make accurate predictions in indoor and outdoor scenes. Moreover, we also found that when the additional supervision of consistency loss is added, the performance is slightly enhanced.

Table 1 Ablation study of our method on NYU-Depth-v2 (NYU), SUN-RGBD (SUN), and Cityscapes (CS) datasets

Data	Improvements			Depth					Segmentation	
	SUM	FSM	$\mathcal{L}_{con}$	higher is better			lower is better		higher is better	
				δ_1	δ_2	δ_3	rms / rms_log	abs_rel / sq_rel	pAcc	mIoU
NYU				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
SUN				0.744	0.932	0.968	0.535 / 0.219	0.168 / 0.134	72.3	39.4
	✓			0.758	0.938	0.973	0.496 / 0.205	0.153 / 0.128	77.1	42.8
		✓		0.755	0.940	0.973	0.501 / 0.211	0.155 / 0.130	75.8	41.6
	✓	✓		0.764	0.944	0.979	0.481 / 0.198	0.149 / 0.125	79.9	44.3
	✓	✓	✓	0.766	0.943	0.979	0.475 / 0.198	0.147 / 0.125	80.7	44.3
CS				0.752	0.882	0.928	7.490 / 0.441	0.251 / 4.159	88.0	63.6
	✓			0.781	0.901	0.943	7.021 / 0.412	0.238 / 3.966	91.5	67.9
		✓		0.776	0.889	0.939	7.124 / 0.427	0.233 / 4.011	90.8	66.5
	✓	✓		0.799	0.907	0.951	6.901 / 0.408	0.227 / 3.878	93.1	69.8
	✓	✓	✓	0.798	0.907	0.951	6.880 / 0.405	0.227 / 3.865	92.9	70.1

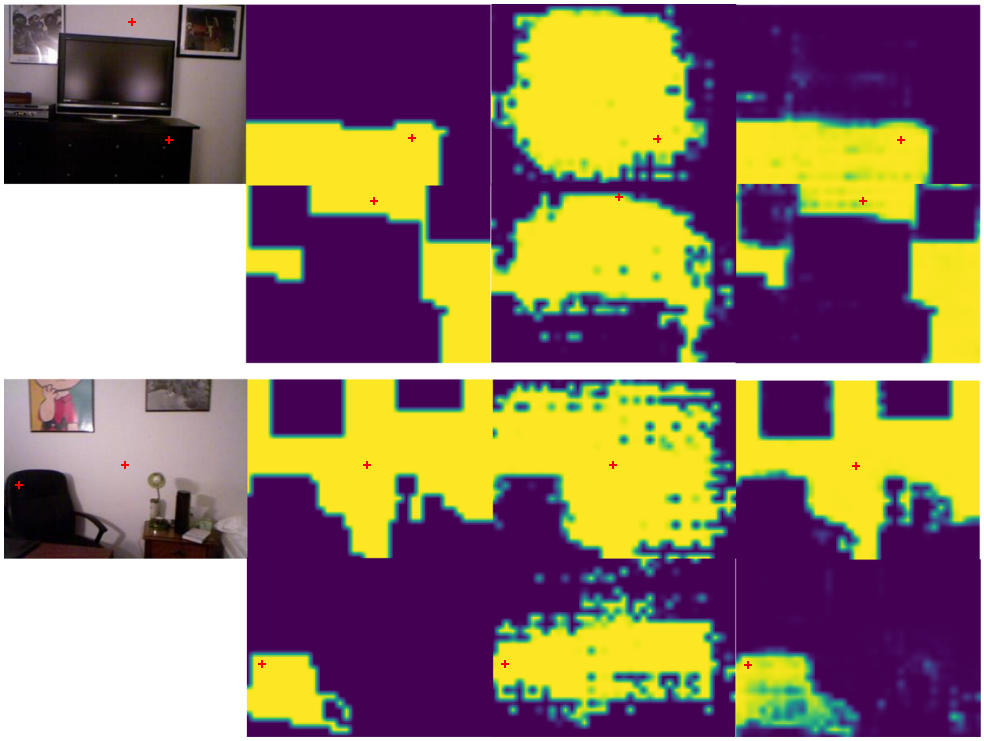

Fig. 7 Comparisons of the learned attention map. **a** input images; **b** ideal attention maps; **c** and **d** represent the attention maps produced by our model without and with supervision of the $\mathcal{L}_{att}$, respectively. For each scene, we show two different attention maps pertaining to the locations where the red plus signs mark

4.4 Comparisons of results

In this subsection, we compare the experimental results of our model with other algorithms according to different datasets.

4.4.1 Results on NYU-Depth-v2

Depth Estimation: We compare the depth estimation results of our approach with some results of representative methods in Table 3. We divide the compared methods into three categories according to the scale of training data, and the signal △ means multi-task learning methods. Among the methods using 795 training pairs, our approach outperforms them on most of the metrics. [42, 44, 45] exploited a single network to predict depth maps. Our method was found to achieve achieves excellent performance. To take the newest published work [45] as a comparison example, although this method obtains excellent values on metrics rms and rms_log, our method outperforms it on δ_1, δ_2, and δ_3. Hence, our designed network, which exploits contextual information and deep task interaction is powerful. Multi-task learning methods [28–30, 43] are also compared; our algorithm is more effective in contrast to other joint learning methods. In addition, we also make comparisons with methods using more data for training. The results show that CI-Net could be on a par with them and even outperforms on some metrics such as rms,verifying the effectiveness of our method. Moreover, we also display some qualitative results in Fig. 9. It can be seen that although the predictions of the method by Laina [17] are smoothed as a whole, they lose some details, bringing in the blurred object boundaries, particularly in the desk, washing machine, and sofa. Besides, the precision of depth maps is weak; the depths of some regions deviate from the ground truth severely. Although [4] has impressive values in evaluation metrics, as seen in Table 3, the contours in predicted depth maps of their models are not sharp, hence, the depth maps are not sufficiently clear. Compared to them,

Table 2 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

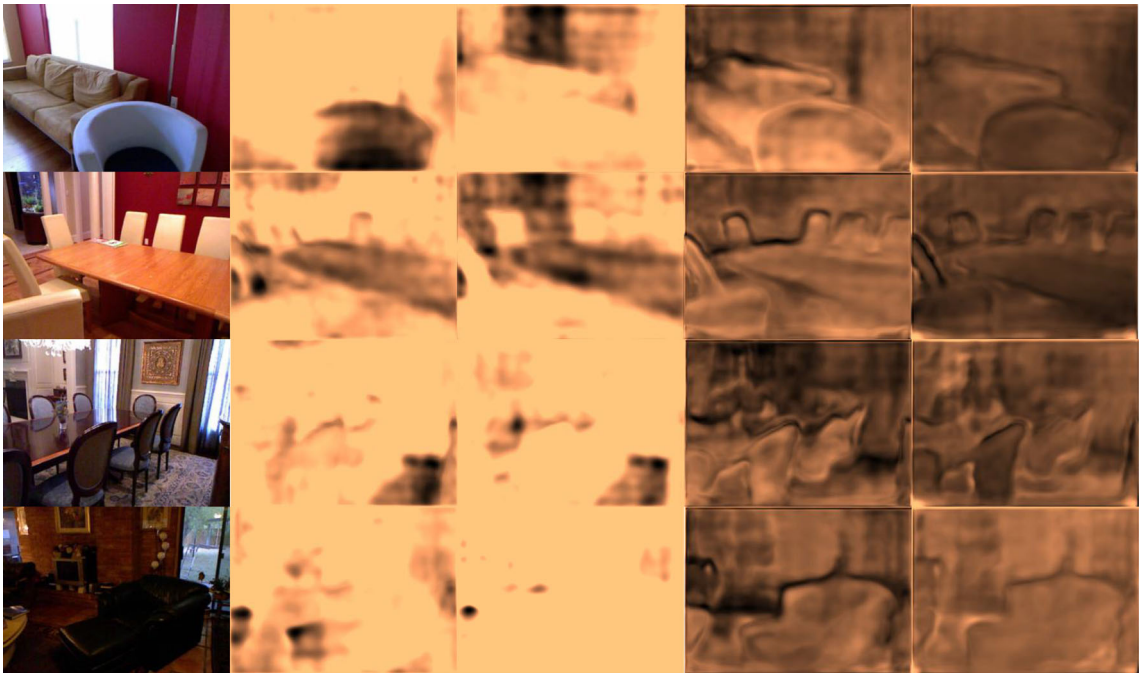

Fig. 8 Visual results of allocated features (summed along the channel dimension and normalized to $0-1$ range). **a** input images; **b** features added to semantic branch learned by LSU; **c** features added to depth branch learned by LSU; **d** features added to semantic branch learned by FSM **e** features added to depth branch learned by FSM

Table 3 Comparisons with the depth estimation methods on NYU-Depth-V2 dataset; △ denotes multi-task learning method

Methods	Scale	higher is better			lower is better			
		δ_1	δ_2	δ_3	rms	rms_log	abs_rel	sq_rel
Roy et al. [42]	795	/	/	/	0.744	/	0.187	/
PAD-Net [29]△	795	0.817	0.954	0.987	0.582	/	0.120	/
C-DCNN [28]△	795	0.736	0.929	0.977	0.628	0.226	0.154	0.136
HybridNet A2 [43]△	795	0.613	0.892	0.974	0.682	0.25	0.202	0.186
Cao et al. [44]	795	0.781	0.954	0.989	0.604	/	0.157	/
SOSD-Net [30]△	795	0.797	0.957	0.991	0.514	/	0.145	/
DP-Net [45]	795	0.784	0.948	0.986	0.474	0.081	/	/
FCRN [17]	12k	0.811	0.953	0.988	0.573	0.195	0.127	/
Li et al. [46]	12k	0.820	0.960	0.989	0.545	/	0.139	/
GeoNet [47]	16k	0.834	0.960	0.990	0.569	/	0.128	/
ACAN [14]	12k	0.815	0.960	0.989	0.518	/	0.144	/
CRFs [4]	95k	0.811	0.954	0.987	0.586	/	0.121	/
DORN [12]	120k	0.828	0.965	0.992	0.509	/	0.115	/
Hu et al. [48]	120k	0.866	0.975	0.993	0.530	/	0.115	/
B-DeNet [49]	95k	/	/	/	0.540	0.183	0.127	/
S2DNet [50]	120k	0.773	0.959	0.989	0.543	/	0.160	/
Cao et al. [44]	120k	0.831	0.962	0.988	0.538	/	0.132	/
CI-Net	795	0.812	0.957	0.990	0.504	0.181	0.129	0.112

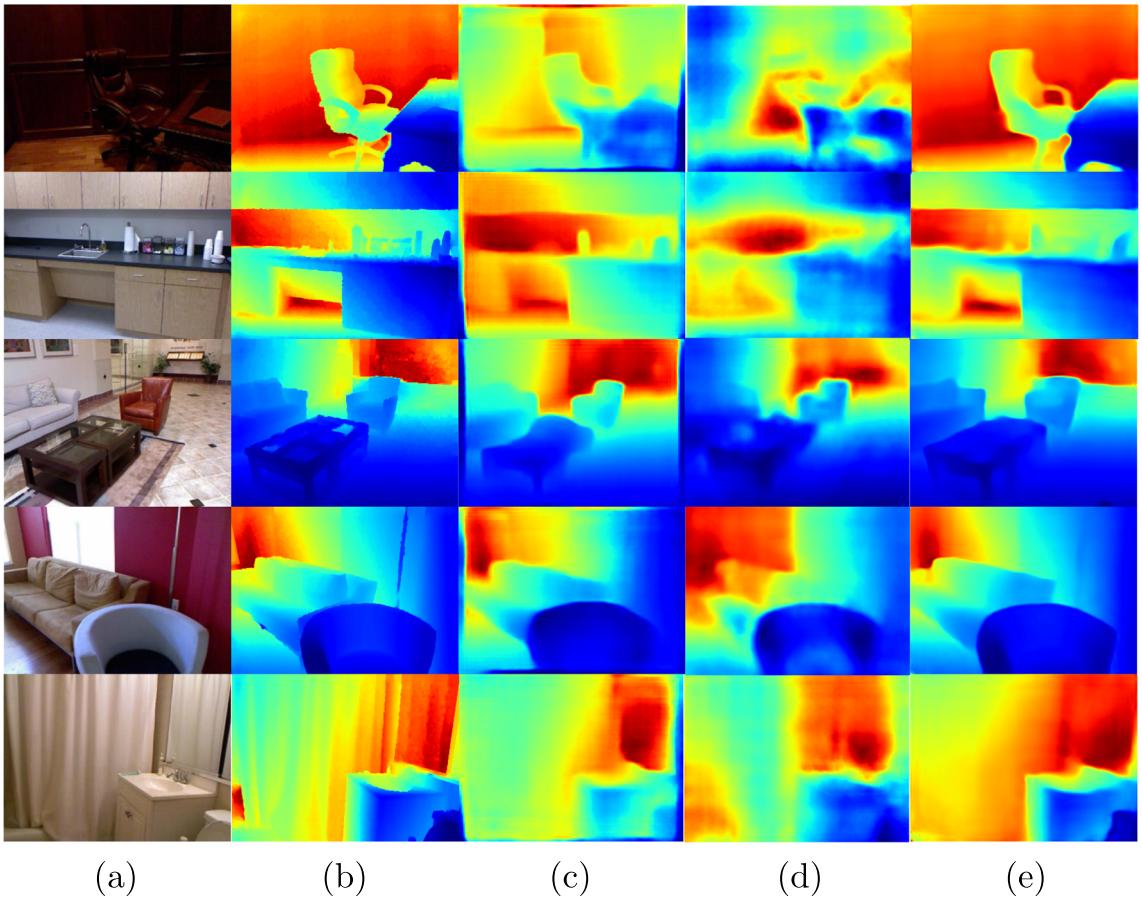

Fig. 9 Visual comparison with some approaches on NYU-depth-v2 dataset. **a** input images; **b** ground truth; **c** predictions of [17]; **d** predictions of [4]; **e** predictions of our model

our results match the structures of scenes and have sharper object boundaries benefiting from prior information of inter and intra-classes.

Semantic Segmentation: Semantic results comparisons are reported in Table 4. According to the types of input images, we classify algorithms into methods using RGB and RGBD images (i.e., using the ground truth depth maps as extra input). It could be noticed that our CI-Net is on a par with all the listed methods in both pixel-acc and mIoU and even performs those multi-task learning methods, demonstrating that the context prior can also benefit the semantic segmentation, and the deep feature interaction does help this task leverage more useful information.

4.4.2 Results on SUN-RGBD

Depth Estimation: As for the SUN-RGBD dataset, we compare our depth estimation results with other methods in Table 5. From the reported values, it can be seen that our method achieves impressive performance on most metrics. Even compared to recent work [50] that uses sparse ground truth depth image as input, our model shows competition on metrics of δ_3 (0.07 improved) and rms (0.208 reduced). The visual results displayed in Fig. 10e further verify that our proposed model is effective in the depth estimation of indoor scenes.

Semantic Segmentation: Quantitative and qualitative results are displayed in Table 4 and Fig. 10, respectively. It could be observed that our method even outperforms those using RGBD images as input, indicating the validity of our proposed FSM and SUM. As for the comparisons with methods using RGB images, although RefineNet-101 achieves best performance in the mIoU metric (45.7), our method is on a par with it on the pixel-acc metric (80.7 vs. 80.4). Some selected visual results are also shown in Fig. 10c; we found that our predictions are of high quality and even give the correct labels of the invalid regions in ground truth.

4.4.3 Results on cityscapes

Depth Estimation: To further evaluate the effectiveness of our algorithm in outdoor scenes, we performed comparison

Table 4 Comparisons with the semantic segmentation methods on NYU-Depth-v2, SUN-RGBD, and Cityscapes datasets for simplicity

Methods	Input	NYU-Depth-v2		SUN-RGBD		Cityscapes	
		pAcc	mIoU	pAcc	mIoU	pAcc	mIoU
FCN [8]	RGB	60.0	29.2	/	/	/	60.0
Context [51]	RGB	70.0	40.6	78.4	42.3	/	/
PSPNet(101) [52]	RGB	72.8	45.2	78.6	44.6	/	79.7
RefineNet-101 [53]	RGB	/	44.7	80.4	45.7	/	/
RefineNet-LW-101 [54]	RGB	/	43.6	/	/	/	/
AdaptNet++ [55]	RGB	/	/	/	38.4	/	81.2
C-DCNN [28]△	RGB	69.0	39.8	77.3	39.0	/	/
SOSD-Net [30]△	RGB	72.2	43.3	/	/	/	68.2
HybridNet A2 [43]△	RGB	71.6	34.3	/	/	93.3	66.6
Ozan et al. [56]△	RGB	/	/	/	/	/	66.6
Cipolla et al. [57]△	RGB	/	/	/	/	/	63.4
SSMA [55]	RGBD	/	/	80.2	43.9	/	/
CMoDE [58]	RGBD	/	/	79.8	41.9	/	/
LFC [59]	RGBD	/	/	79.4	41.8	/	/
FuseNet [23]	RGBD	/	/	76.3	37.3	/	/
D-CNN [60]	RGBD	/	41.0	/	42.0	/	/
CI-Net	RGB	72.7	42.6	80.7	44.3	92.9	70.1

△ denotes multi-task learning method. We use NYU, SUN, and CS to represent NYU-Depth-v2, SUN-RGBD, and Cityscapes

experiments on the Cityscapes dataset. As observed in Table 5, compared with the methods [11, 29, 43], the presented CI-Net achieves a significant improvement, indicating that our method is more powerful than the listed joint-task learning methods in outdoor scenes. Furthermore, our model is competitive with the recently proposed work SDC-Depth [61] and even obtains better values on the metrics of rms and rms_log. We also display some predicted depth maps of our method in Fig. 11e. From the results, we could find that the depth maps predicted by our method have sharp boundaries around objects, and the depth values of objects such as cars, ground, and buildings are accurately predicted.

Semantic Segmentation: Meanwhile, we also compare the semantic results of our network on the Cityscapes dataset with semantic segmentation methods. In contrast to joint-task

Table 5 Comparisons with the depth estimation methods on SUN-RGBD and Cityscapes dataset

Methods	Data	higher is better			lower is better			
		δ_1	δ_2	δ_3	rms	rms_log	abs_rel	sq_rel
Cao et al. [5]	SUN	0.563	0.727	0.882	0.839	/	0.256	/
C-DCNN [28]△		0.792	0.948	0.985	0.538	0.215	0.159	0.108
Zhang et al. [11]△		/	/	/	0.468	/	0.140	/
S2DNet [50]□		0.881	0.951	0.972	0.683	/	0.122	/
CI-Net	SUN	0.766	0.943	0.979	0.475	0.198	0.147	0.125
FCRN [17]	CS	0.765	0.893	0.940	7.273	0.448	0.257	4.238
PAD-Net [29]△		0.786	0.905	0.945	7.117	0.428	0.246	4.060
HybridNet A2 [43]△		0.748	0.822	0.929	12.09	0.434	0.240	4.27
Zhang et al. [11]△		0.776	0.903	0.945	7.104	0.416	0.234	3.776
SDC-Depth [61]		0.801	0.913	0.950	6.917	0.414	0.227	3.800
CI-Net	CS	0.798	0.907	0.951	6.880	0.405	0.227	3.865

Note that SUN denotes images from SUN-RGBD, and CS indicates the dataset of Cityscapes. The signal △ denotes multi-task learning method, and □ represents that this method uses sparse depth maps as extra input

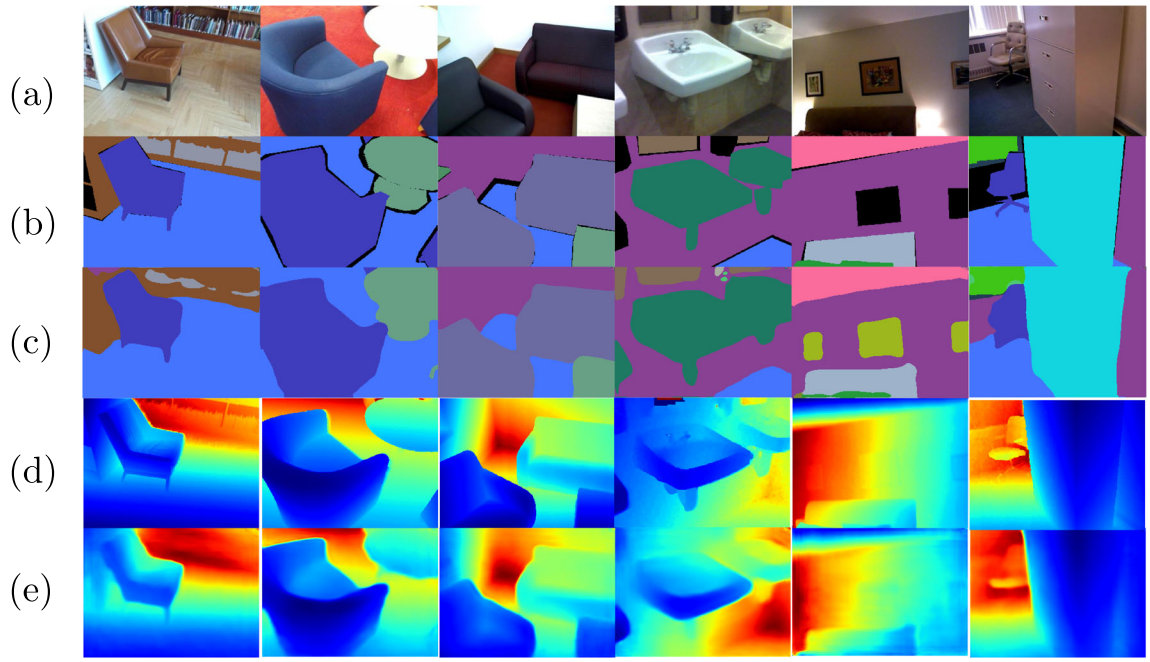

Fig. 10 Visualized depth and semantic segmentation maps on SUN-RGBD dataset. **a** input images; **b** ground truth of semantic segmentation; **c** semantic segmentation predictions of our method; **d** ground truth of depth maps; **e** depth estimation of our method

learning methods [30, 43, 56, 57], our approach achieves best results on pixel-acc and mIoU, and it could be noticed that our method outperforms them on the mIoU metric (2.7% improved compared with newly proposed method SOSD-Net). Some visual results are also displayed in Fig. 11c, where it can be observed that buildings, cars, sky, and trees are correctly predicted, demonstrating that our method is suitable for outdoor scenes too.

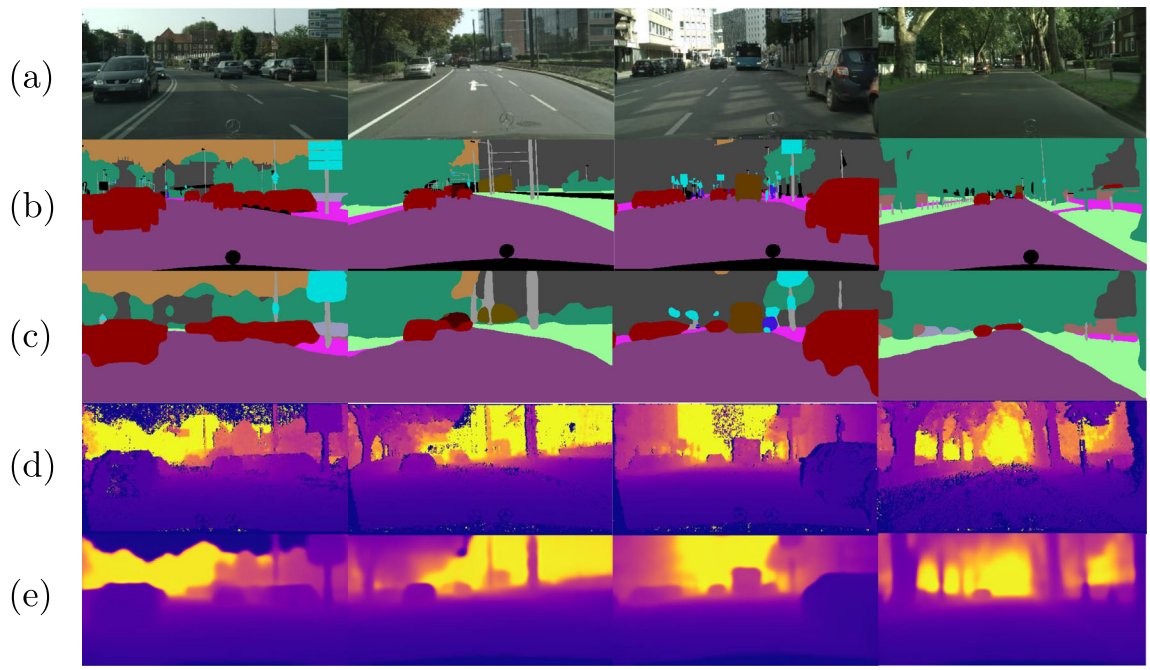

Fig. 11 Visualized depth and semantic segmentation maps on Cityscapes dataset. **a** input images; **b** ground truth of semantic segmentation; **c** semantic segmentation predictions of our method; **d** ground truth of depth maps; **e** depth estimation of our method

5 Conclusion

In this paper, a network for joint-task learning was proposed. By employing the scene-understanding module, the presented network was able to capture the contextual information of inter and intra-classes, which is crucial for the network to understand which useful context can be exploited to make predictions. To fuse the task-specific features deeply, we designed an FSM that enlarged the receptive fields and augmented the presentation ability through upsampling and downsampling operations. In addition, a consistency loss was proposed to make the task-specific features mutually guided, which kept consistent relationships with the respective adjacent features. The performance of ablation study and the comparative results with other methods on NYU-Depth-v2, SUN-RGBD, and Cityscapes datasets demonstrated the effectiveness of our method. 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 and edge detection, into this work. We are also interested in combining the depth prediction with semantic SLAM to obtain more accurate results.

Declarations

Conflict of Interests (check journal-specific guidelines for which heading to use): Not Available

References

1. Fang B, Mei G, Yuan X, Wang L, Wang Z, Wang J (2021) Visual slam for robot navigation in healthcare facility. Pattern Recogn 113:107822. https://doi.org/10.1016/j.patcog.2021.107822
2. Husbands P, Shim Y, Garvie M, Dewar A, Domcsek N, Graham P, Knight J, Nowotny T, Philippides A (2021) Recent advances in evolutionary and bio-inspired adaptive robotics: Exploiting embodied dynamics. Appl Intell 51(9):6467–6496. https://doi.org/10.1007/s10489-021-02275-9
3. Lee D-H, Chen K-L, Liou K-H, Liu C-L, Liu J-L (2020) Deep learning and control algorithms of direct perception for autonomous driving. Appl Intell 51(1):237–247. https://doi.org/10.1007/s10489-020-01827-9
4. Xu D, Wang W, Tang H, Liu H, Sebe N, Ricci E (2018) Structured attention guided convolutional neural fields for monocular depth estimation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition
5. Cao Y, Wu Z, Shen C (2018) Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Trans Circ Syst Video Technol 28(11):3174–3182. https://doi.org/10.1109/tcsvt.2017.2740321
6. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: 2018 IEEE/CVF conference on computer vision and pattern recognition
7. Lan X, Gu X, Gu X (2021) MMNet: Multi-modal multi-stage network for RGB-t image semantic segmentation. Appl Intell. https://doi.org/10.1007/s10489-021-02687-7
8. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
9. Guizilini V, Hou R, Li J, Ambrus R, Gaidon A (2019) Semantically-guided representation learning for self-supervised monocular depth. In: International conference on learning representations
10. Zhang Z, Cui Z, Xu C, Yan Y, Sebe N, Yang J (2019) Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
11. Zhang Z, Cui Z, Xu C, Jie Z, Li X, Yang J (2020) Joint task-recursive learning for rgb-d scene understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2608–2623. https://doi.org/10.1109/TPAMI.2019.2926728
12. Fu H, Gong M, Wang C, Batmanghelich K, Tao D (2018) Deep ordinal regression network for monocular depth estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition
13. Jiao J, Cao Y, Song Y, Lau R (2018) Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In: Computer Vision – ECCV 2018, pp 55–71
14. Chen Y, Zhao H, Hu Z, Peng J (2021) Attention-based context aggregation network for monocular depth estimation. Int J Mach Learn Cybern 12(6):1583–1596. https://doi.org/10.1007/s13042-020-01251-y
15. Yu C, Wang J, Gao C, Yu G, Shen C, Sang N (2020) Context prior for scene segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
16. Klingner M, Termöhlen J-A, Mikolajczyk J, Fingscheidt T (2020) Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: Computer Vision – ECCV 2020, pp 582–600
17. Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D Vision (3DV)
18. Yin W, Liu Y, Shen C (2021) Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. IEEE Trans Pattern Anal Mach Intell:1–1. https://doi.org/10.1109/TPAMI.2021.3097396
19. Zhou W, Zhou E, Liu G, Lin L, Lumsdaine A (2020) Unsupervised monocular depth estimation from light field image. IEEE Trans Image Process 29:1606–1617. https://doi.org/10.1109/TIP.2019.2944343
20. Ye X, Fan X, Zhang M, Xu R, Zhong W (2021) Unsupervised monocular depth estimation via recursive stereo distillation. IEEE Trans Image Process 30:4492–4504. https://doi.org/10.1109/TIP.2021.3072215
21. Wu Y, Jiang J, Huang Z, Tian Y (2021) Fpanet: Feature pyramid aggregation network for real-time semantic segmentation. Appl Intell:1–18. https://doi.org/10.1007/s10489-021-02603-z
22. Qi X, Liao R, Jia J, Fidler S, Urtasun R (2017) 3d graph neural networks for RGBD semantic segmentation. In: 2017 IEEE International Conference on Computer Vision (ICCV)
23. Hazirbas C, Ma L, Domokos C, Cremers D (2017) FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Computer Vision – ACCV 2016, pp 213–228
24. Sun L, Yang K, Hu X, Hu W, Wang K (2020) Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images. IEEE Robot Autom Lett 5(4):5558–5565. https://doi.org/10.1109/LRA.2020.3007457
25. Hu X, Yang K, Fei L, Wang K (2019) ACNET: Attention based network to exploit complementary features for RGBD semantic

segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP)
26. Hung S-W, Lo S-Y, Hang H-M (2019) Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP)
27. Chen L-Z, Lin Z, Wang Z, Yang Y-L, Cheng M-M (2021) Spatial information guided convolution for real-time RGBD semantic segmentation. IEEE Trans Image Process 30:2313–2324. https://doi.org/10.1109/tip.2021.3049332
28. Liu J, Wang Y, Li Y, Fu J, Li J, Lu H (2018) Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. IEEE Trans Neural Netw Learn Syst 29(11):5655–5666. https://doi.org/10.1109/TNNLS.2017.2787781
29. Xu D, Ouyang W, Wang X, Sebe N (2018) PAD-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: 2018 IEEE/CVF Conference on computer vision and pattern recognition
30. He L, Lu J, Wang G, Song S, Zhou J (2021) SOSD-net: Joint semantic object segmentation and depth estimation from monocular images. Neurocomputing 440:251–263. https://doi.org/10.1016/j.neucom.2021.01.126
31. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
32. Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, Huijben I, Chennakeshava N, Mento F, Sentelli A, Peschiera E, Trevisan R, Maschietto G, Torri E, Inchingolo R, Smargiassi A, Soldati G, Rota P, Passerini A, van Sloun RJG, Ricci E, Demi L (2020) Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 39(8):2676–2687. https://doi.org/10.1109/TMI.2020.2994459
33. Chen T, An S, Zhang Y, Ma C, Wang H, Guo X, Zheng W (2020) Improving monocular depth estimation by leveraging structural awareness and complementary datasets. In: Computer Vision – ECCV 2020, pp 90–108
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
35. Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
36. Huynh L, Nguyen-Ha P, Matas J, Rahtu E, Heikkilä J (2020) Guiding monocular depth estimation using depth-attention volume. In: Computer Vision – ECCV 2020, pp 581–597
37. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: Computer Vision – ECCV 2012, pp 746–760
38. Song S, Lichtenberg SP, Xiao J (2015) SUN RGB-d: A RGB-d scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
39. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
40. Ming Y, Meng X, Fan C, Yu H (2021) Deep learning for monocular depth estimation: a review. Neurocomputing 438:14–33. https://doi.org/10.1016/j.neucom.2020.12.089
41. Mohammadi Amiri M, Gündüz D (2020) Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. IEEE Trans Signal Process 68:2155–2169. https://doi.org/10.1109/TSP.2020.2981904
42. Roy A, Todorovic S (2016) Monocular depth estimation using neural regression forest. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 5506–5514
43. Lin X, Sánchez-Escobedo D, Casas JR, Pardàs M (2019) Depth estimation and semantic segmentation from a single rgb image using a hybrid convolutional neural network. Sensors 19(8). https://doi.org/10.3390/s19081795
44. Cao Y, Zhao T, Xian K, Shen C, Cao Z, Xu S (2020) Monocular depth estimation with augmented ordinal depth relationships. IEEE Trans Circ Syst Video Technol 30(8):2674–2682. https://doi.org/10.1109/TCSVT.2019.2929202
45. Ye X, Chen S, Xu R (2021) Dpnet: Detail-preserving network for high quality monocular depth estimation. Pattern Recogn 109:107578. https://doi.org/10.1016/j.patcog.2020.107578
46. Li B, Dai Y, He M (2018) Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference. Pattern Recogn 83:328–339. https://doi.org/10.1016/j.patcog.2018.05.029
47. Qi X, Liao R, Liu Z, Urtasun R, Jia J (2018) GeoNet: Geometric neural network for joint depth and surface normal estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition
48. Hu J, Ozay M, Zhang Y, Okatani T (2019) Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)
49. Yang X, Gao Y, Luo H, Liao C, Cheng K-T (2019) Bayesian denet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty. IEEE Trans Multimed 21(11):2701–2713. https://doi.org/10.1109/TMM.2019.2912121
50. Hambarde P, Murala S (2020) S2dnet: Depth estimation from single image and sparse samples. IEEE Trans Comput Imaging 6:806–817. https://doi.org/10.1109/TCI.2020.2981761
51. Lin G, Shen C, van den Hengel A, Reid I (2016) Efficient piecewise training of deep structured models for semantic segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3194–3203
52. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
53. Lin G, Milan A, Shen C, Reid I (2017) RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
54. Nekrasov V, Shen C, Reid I (2018) Light-weight refinenet for real-time semantic segmentation. In: Proceedings of the british machine vision conference, pp 278–284
55. Valada A, Mohan R, Burgard W (2019) Self-supervised model adaptation for multimodal semantic segmentation. Int J Comput Vis 128(5):1239–1285. https://doi.org/10.1007/s11263-019-01188-y
56. Sener O, Koltun V (2018) Multi-task learning as multi-objective optimization. In: NeurIPS
57. Cipolla R, Gal Y, Kendall A (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition
58. Valada A, Vertens J, Dhall A, Burgard W (2017) AdapNet: Adaptive semantic segmentation in adverse environmental conditions. In: 2017 IEEE International Conference on Robotics and Automation (ICRA)
59. Valada A, Oliveira GL, Brox T, Burgard W (2017) Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In: Springer Proceedings in Advanced Robotics, pp 465–477
60. Wang W, Neumann U (2018) Depth-aware CNN for RGB-d segmentation. In: Computer Vision – ECCV 2018, pp 144–161

61. Wang L, Zhang J, Wang O, Lin Z, Lu H (2020) SDC-depth: Semantic divide-and-conquer network for monocular depth estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Tianxiao Gao received his bachelor degree in Beijing Forestry University, Beijing, China. He is currently pursuing the master degree from the School of Automation Science and Engineering, South China University of Technology. His research focus on monocular depth estimation, semantic segmentation and visual SLAM.

Wu Wei received a Ph.D. degree from Huazhong University of Science and Technology in 2000. From January 2001 to December 2002, he was a postdoctoral researcher at the Department of Automation at Tsinghua University. From 2003 to 2004, he worked as a researcher at the Institute of Intelligent Systems, The Hong Kong Polytechnic University. He is also a member of the International IEEE Control Systems, Automotive Technology, Signal Processing, System People and Control Systems, Robotics and Automation Association. He is currently as a professor at the School of Automation Science and Engineering, South China University of Technology. His research focus on intelligent control, intelligent systems, robot control technology, pattern recognition and artificial intelligence.

Zhongbin Cai received his bachelor degree in South China University of Technology, Guangzhou, China. He is currently pursuing the master degree from the School of Automation Science and Engineering, South China University of Technology. His research focus on reinforcement learning, robotic control and deep learning.

Fan Zhun received a Ph. D. degree from electrical engineering, Michigan State University, USA, in 2004. From 2004 to 2007, he worked as an assistant professor at the Technical University of Denmark, Denmark. From 2007 to 2011, he was an associate professor at the Department of Mechanical Engineering and the Department of Management Engineering at the Technical University of Denmark. He is currently as a professor at the College of Engineering, Shantou University. His research interest covers artificial intelligence, mechatronics design automation, intelligent robot system, computer vision, and evolutionary computation.

Sheng Quan Xie received the Ph.D. degree in mechanical engineering from Huazhong University of Science and Technology, Wuhan, China, in 1998, and the Ph.D. degree in mechanical engineering from the University of Canterbury, Christchurch, New Zealand, in 2002. From 2003 to 2016, he was with the University of Auckland, Auckland, New Zealand, where he chaired the research group of Biomechatronics. Since 2017, he has been with the University of Leeds. He has published 5 books, 15 book chapters, and more than 280 academic papers. His research interests include medical and rehabilitation robots and advanced robot control. Prof. Xie is a Fellow of The Institution of Professional Engineers New Zealand. He has also served as a Technical Editor of the IEEE/ASME TRANSACTIONS ON MECHATRONICS.

Xinmei Wang received her B.S. and M.S. degrees from Wuhan University of Technology, and Ph.D. degree from South China University of Technology in 2009. She is currently a lecturer at School of Automation, China University of Geosciences. Her research interests include stability analysis and control of timedelay system, robot visual servoing and switching system.

Qiuda Yu received his bachelor degree in Hangzhou Dianzi University, Zhejiang, China. He is currently pursuing the Ph.D. from the School of Automation Science and Engineering, South China University of Technology. His research focus on visual grasp, visual guided control and deep-learning.

Affiliations

Tianxiao Gao[1] · Wu Wei[1] · Zhongbin Cai[1] · Zhun Fan[2] · Sheng Quan Xie[3] · Xinmei Wang[4] · Qiuda Yu[1]

✉ Zhun Fan
zfan@stu.edu.cn

Tianxiao Gao
201920116403@scut.edu.cn

Zhongbin Cai
201921017250@scut.edu.cn

Sheng Quan Xie
s.q.xie@leeds.ac.uk

Xinmei Wang
wangxm@cug.edu.cn

Qiuda Yu
yuqiuda@163.com

1 School of Automation Science and Engineering, South China University of Technology, Guangzhou, 510000, Guangdong, China

2 College of Engineering, Shantou University, Shantou, 515000, Guangdong, China

3 School of Electronic and Electrical Engineering, University of Leeds, Leeds, LS2 9JT, UK, China

4 School of Automation, China University of Geosciences, Wuhan, 430074, Hubei, China