Scale-aware Network with Modality-awareness for RGB-D Indoor Semantic Segmentation

Feng Zhou[a,*], Yu-Kun Lai[b], Paul L. Rosin[b], Fengquan Zhang[a], Yong Hu[c,d]

[a]*North China University of Technology*
[b]*Cardiff University*
[c]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University*
[d]*School of New Media Art and Design, Beihang University*

Abstract

This paper focuses on indoor semantic segmentation based on RGB-D images. Semantic segmentation is a pixel-level classification task that has made steady progress based on fully convolutional networks (FCNs). However, we find there is still room for improvements in the following three aspects. The first is related to multi-scale feature extraction. Recent state-of-the-art works forcibly concatenate multi-scale feature representations extracted by spatial pyramid pooling, dilated convolution or other architectures, regardless of the spatial extent for each pixel. The second is regarding RGB-D modal fusion. Most successful methods treat RGB and depth as two separate modalities and force them to be joined together regardless of their different contributions to the final prediction. The final aspect is about the modeling ability of extracted features. Due to the "local grid" defined by the receptive field, the learned feature representation lacks the ability to model spatial dependencies. In addition to these modules, we design a depth estimation module to encourage the RGB network to extract more effective features. To solve the above challenges, we propose four modules to address them: scale-aware module, modality-aware module, attention module and depth estimation module. Extensive experiments on the NYU-Depth v2 and SUN RGB-D datasets demonstrate that our method is effective against RGB-D indoor semantic segmentation.

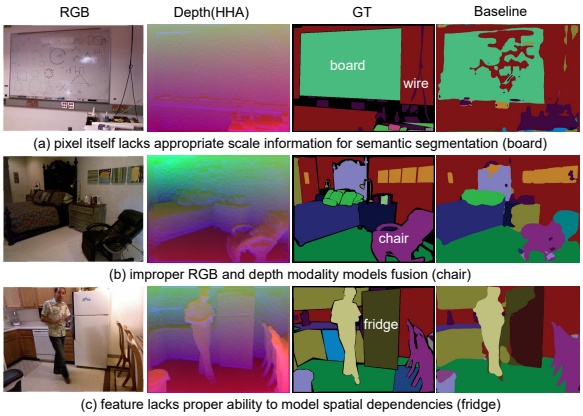
Keywords: Semantic segmentation, Scale selection, Attention, RGB-D, Depth estimation

1. Introduction

The purpose of semantic segmentation is to assign specific class labels to regions in the input images. This is a fundamental task for scene understanding [1], video analysis [1, 2], clothing retrieval [3], and such of those intelligent applications [4]. However, scene understanding is a daunting task, especially for indoor scenes, due to the varying illuminations and cluttered backgrounds. With the development of commercial depth cameras, such as Kinect and Prime-Sense, we are able to capture high-quality, synchronized RGB and depth images. RGB data provides rich visual information such as color and texture. In contrast to RGB data, the depth modality data provides pure shape and geometry information, which is invariant to lighting and reflectance. Combining these two complementary modalities provides us with an opportunity to dramatically improve the performance of semantic segmentation of indoor scenes.

Extensive studies have been conducted for the task of indoor semantic segmentation. [5] proposes a patch-wise model, and [6] utilizes an R-CNN (Region-based Convolutional Neural Network) scheme to learn an RGB-D multi-modal feature representation to boost the performance. Recently, [7] proposes an end-to-end FCN (Fully Convolutional Network) for semantic segmentation and achieves significant improvement. However, there are still many problems with indoor semantic segmentation. Towards the problem of multi-scale objects, many successful methods [8, 9, 10, 11, 12] adopt pyramid layers to extract multi-scale feature representations. Towards the problem of modeling long-range contextual information, [9, 13, 14] utilize global pooling techniques to obtain global context feature, and [15] subdivides images into super-pixels and uses LSTM

*Corresponding author
Email addresses: zhoufeng@ncut.edu.cn (Feng Zhou), laiy4@cardiff.ac.uk (Yu-Kun Lai), RosinPL@cardiff.ac.uk (Paul L. Rosin), fqzhang@ncut.edu.cn (Fengquan Zhang), huyong@buaa.edu.cn (Yong Hu)



(a) pixel itself lacks appropriate scale information for semantic segmentation (board)

(b) improper RGB and depth modality models fusion (chair)

(c) feature lacks proper ability to model spatial dependencies (fridge)

Figure 1: Limitations of the baseline on indoor scene semantic segmentation with RGB-D data. The depth image in this paper is encoded to three channel HHA (horizontal disparity, height above ground, and angle with gravity). The baseline consists of two-stream atrous spatial pyramid pooling networks trained on RGB and depth data respectively. These two networks are combined together by late fusion with equal-weight sum.

(Long Short-Term Memory) to aggregate and enlarge contextual information by multi-scale context intertwining. Towards RGB-D fusion, three levels of fusion are often adopted. The first one is early fusion [5], which simply concatenates the input of two complementary modalities, RGB and depth, together as four-channel input. The second one is middle fusion [6], which leverages the two modalities, RGB and depth, as two independent inputs and extracts different modality feature representations, and then concatenates them together to learn a final classifier. The third one is late fusion (also called score map fusion) [7], which utilizes RGB and depth as two separate inputs to learn two different models, and obtain two different score maps. Then, the two score maps are fused together by equal weights.

In this paper, the model proposed by [8] is extended by using the late fusion strategy. This extended model is used as our baseline for indoor semantic segmentation in this study. Compared with [7], our extended model achieves better performance. However, we have found that there are three aspects that can be improved. The first is that the pixel itself does not have enough information for semantic prediction, it needs to learn the appropriate scale information. As shown in Figure 1(a), since the appearance of the board object is very similar to the back wall, multi-scale feature representations extracted using multiple atrous convolutional layers do not provide proper surrounding scale information for pixels on the board. Likewise, for the wire object, since it is so thin the extracted multi-scale features do not capture suitable information for it. The second is about the fusion of two complementary modalities. As shown in Figure 1(b), the appearance cues are beneficial for classifying the object as a chair, whereas the depth cue would confuse the identification of chair parts (most sofa objects in the dataset are near a wall). The third one is that the extracted features lack the ability to model long-range dependencies. As shown in Figure 1(c), part of the refrigerator is misclassified as a door (due to being confused by the rectangle shape).

This paper aims to discuss the problem of indoor semantic segmentation based on the two complementary modalities of RGB and depth. In particular, we propose a scale-aware module, a modality-aware module, and an attention module, which address the above three-aspect problems. The scale-aware module learns a proper scale feature representation for each object in the input. It learns a weighted mask for each extracted multi-scale feature, and then multiplies these masks by the multi-scale features to generate a scale-aware feature representation to address the first problem. Towards the second problem, the modality-aware module is proposed to combine the two complementary modalities of RGB and depth using different weights instead of equal weights. Towards the third problem, an attention module is introduced to complement the scale-aware module, which can capture long-range dependencies in the generated scale-aware feature representation. Besides the above three modules, since we have the ground truth depth value of the input RGB image, we can thus design an encoder-decoder depth estimation module on the RGB network to encourage the RGB backbone network to extract better and more precise features. The contributions of this paper can be summarized in the following aspects.

- An efficient scale-aware module with modality-awareness, an attention module, and a depth estimation network is proposed for semantic segmentation.
- Within the network, a scale-aware module is used to select the appropriate scale feature for each pixel, which enables a proper scale feature representation to be learned for each object in the input.
- In order to improve the segmentation performance, a modality-aware module is proposed, which adaptively combines the RGB module and depth module to obtain useful features.
- To further improve the segmentation performance, the attention module and depth estimation module

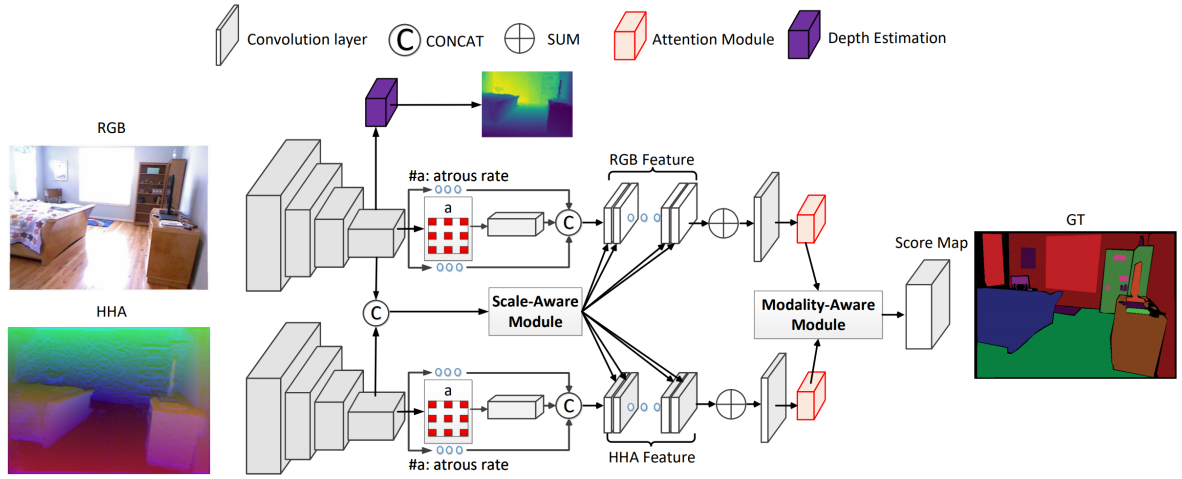


Figure 2: The overall architecture of our SAMD model for RGB-D indoor semantic segmentation. It is a two-stream convolutional neural network, one for RGB and the other one for depth (HHA). SAMD consists of four parts: 1) the encoder feature extractor part. It is a standard two-stream convolutional neural network, which leverages atrous spatial pyramid pooling to learn multi-scale feature representations; 2) the scale-aware module, which is used to learn features maps of an appropriate scale; 3) the modality-aware module, which is proposed to effectively combine RGB and depth networks based on the contributions of the two modalities; 4) the attention and depth estimation module, which is used to extract more plausible features. Best viewed in color.

are proposed to extract better feature representations. The former is to obtain long-range dependent features, and the latter is to force the RGB module to extract more plausible feature representations.

The rest of the paper is organized as follows. Section 2 briefly covers related work, highlighting current work. Then we give the details of the proposed approach in Section 3. Experimental results and analysis are provided in Section 4. Finally, the conclusions are drawn in Section 5.

2. Related Work

The proposed method relates to a lot of work on scale-aware selection, attention method, modal combination and depth estimation. CNN-based semantic segmentation has achieved great advances in recent years [16, 8, 13, 9, 17, 18, 19, 20]. Most of the existing work has employed fully convolutional networks (FCNs) [7]. However, objects in indoor scenes cover a huge range of scales due to both their range of actual sizes in the real world, as well as by their differences in distance to the camera. The methods above only forcibly stack the extracted multi-scale features together. This is not enough for real-world cluttered indoor scene understanding.

Selecting the appropriate scale feature for each pixel is particularly important. Many successful works have investigated this problem. [21] proposes a channel attention scheme to boost the performance of semantic segmentation. [22] exploits a scale-space to select a properly scaled feature. However, as far as we know, there is little work on RGB-D feature scale selection. In this paper, we propose a scale-aware module that combines RGB and depth modal features to build a scale-aware module to improve the performance of RGB-D semantic segmentation. Although the scale-aware module can generate features that fit the scale for each neuron, the feature cannot reflect the contributions of each modality.

The synchronized RGB and depth pair images provide useful multi-modal information for the task of computer vision. Most successful methods simply combine the extracted multi-modal feature representation using early fusion [5], middle fusion [6], or late fusion [7]. However, in the final prediction layer, RGB and depth contribute unequally in most cases. An example is shown in Figure 1 (b) where the chair object is misclassified by concatenating the two complementary modalities with the same weight.

Recently, the attention mechanism has been proposed to model and capture long-range dependencies, and it has become an integral part of many successful works [23, 24, 25, 26]. [27] proposes a self-attention mechanism to capture long-range dependencies of inputs and achieves the state-of-the-art performance in machine translation. The attention mechanism has not only been

used in the Natural Language Processing (NLP) field, but has also been utilized in the computer vision field. [28] utilizes a self-attention scheme to obtain better performance on the image generation task. [29] adopts an attention mechanism in object recognition to boost performance. [30] proposes a MAT (Motion-Attentive Transition) module comprised of a soft attention unit and an attention transition unit to learn more specific and useful feature representations.

The combination of semantic segmentation and depth estimation was studied in many previous works, with the goal of improving both semantic segmentation and depth estimation. [31] proposes three ways to improve semantic segmentation performance with depth estimation, and [32] adopts knowledge from a semantic segmentation network to teach the depth estimation task. In our paper, we introduce depth estimation as an auxiliary task to help improve semantic segmentation.

This paper adopts a scale-aware module, a modality-aware module, a self-attention and a depth estimation module to address the above problems. As shown in the experiments, the proposed four modules can achieve performance gains on many publicly RGB-D semantic segmentation datasets.

3. Our Approach

In the following section, we mainly focus on the learning details of the proposed SAMD method. SAMD is composed of four modules: the scale-aware module, the modality-aware module, attention, and depth estimation (as shown in Figure 2). The scale-aware module is used to generate a scale-aware feature representation which predicts the scale information for each pixel from the learned multi-scale feature representation. The modality-aware module is to learn an effective fusion way for the two modal networks. To further improve the performance, we propose the attention and depth estimation modules. The attention module is used to capture the global feature dependencies in the spatial domain for the input feature. The depth estimation module is used to push the RGB network to extract more precise and useful features.

We adopt atrous spatial pyramid pooling (ASPP) as our feature encoder to extract multi-scale features. To be specific, let $\mathcal{L} = \{(\mathcal{R}_1, \mathcal{D}_1, Y_1), ..., (\mathcal{R}_n, \mathcal{D}_n, Y_n)\}$ be the n pairwise RGB-D training data, where $\mathcal{R} = \{r_i\}_{i=1}^{H\times W}$ is the RGB modality training image whose size is $H \times W$, and $\mathcal{D} = \{d_i\}_{i=1}^{H\times W}$ is the corresponding depth training image, whose size is $H \times W$, and $Y = \{y_i\}_{i=1}^{H\times W}$ is the label image, in which r_i and d_i are corresponding pixels in the pairwise image, label $y_i \in \{0, 1, ..., C\}$ gives the per-pixel label, C denotes the number of the categories. In our approach, given an $H \times W$ pair RGB-D image, through the encoder part, we obtain features f_e^r and f_e^d whose sizes are $\frac{H}{8} \times \frac{W}{8}$ (ignoring the channel size), where f_e^r is from RGB modality, and f_e^d is from depth modality. These two features serve two purposes. The first is to generate subsequent multi-scale feature representations and the second is used in our scale-aware module for scale selection.

3.1. Scale-aware Module

The output of the feature encoder part is a multi-scale feature of the forced concatenation, but the learned feature still does not hold the correct scale feature representation. To this end, we employ a scale-aware module to enable our model to learn a feature map of proper scale for all neurons in the input.

Specifically, let the multi-scale feature set generated from f_e^r and f_e^d be $\{f_{a_i}^r, f_{a_i}^d\}$, where $f_{a_i}^r$ denotes the multi-scale feature extracted from the RGB modality feature encoder part by dilated convolution (a.k.a. atrous convolution) [8] (kernel size a_i), $f_{a_i}^d$ denotes the feature from the depth modality. In the experiments, we adopt four dilated kernel sizes (6, 12, 18, 24) for each modality, and a_i stands for the 4 different dilated kernels. We concatenate f_e^r and f_e^d to generate f_e^f, and feed it into a 1×1 convolutional layer $conv(.)$ and output f_c^f whose size is $8 \times \frac{H}{8} \times \frac{W}{8}$. Then we use a softmax operation to normalize f_c^f to obtain f_m^f. For the RGB modality, we split the four channel feature representation (corresponding to the four channels of f_e^r) $f_{m_j}^f$ in f_m^f and then calculate the scale-aware features f_{sa}^r and f_{sa}^d as follows:

$$f_{sa}^r = \sum_{j=1}^{4} f_{a_i}^r \odot f_{m_j}^f \tag{1}$$

to the depth modality,

$$f_{sa}^d = \sum_{j=5}^{8} f_{a_i}^d \odot f_{m_j}^f \tag{2}$$

where the operator $\odot$ represents the Hadamard product.

3.2. Modality-aware Module

The modality-aware module is proposed to combine feature representations of RGB and depth modalities for semantic segmentation. The structure of the module is

Figure 3: Illustration of the scale-aware confidence map. The two images in the first column are RGB and depth images. The two images in the second column are scale-aware confidence maps upon the two modalities. The remaining 4 images in the first row are each scale channel confidence map of RGB modality, in the second row are each channel confidence map of depth modality. From left to right, they are a_6, a_{12}, a_{18}, a_{24}. For the sake of simplicity, we show the confidence maps (average value) by using the "COLORMAP_JET" color map (where blue is low value and red is high value) upon RGB. Best viewed in color.

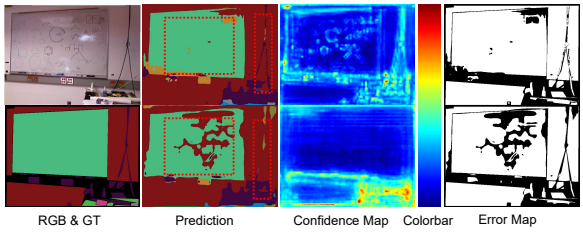


Figure 4: Illustration of the scale-aware module. From the illustration, we can find that with the scale-aware module, our model can effectively focus on larger and smaller objects, as shown by the red dashed boxes. From left to right and top to bottom, they are RGB and ground truth; prediction results with and without the scale-aware module; confidence maps with and without scale-aware module (including the colorbar where the value increases from blue to red); error maps with and without scale-aware module. Best viewed in color.

similar to the scale-aware module and it is composed of four layers. The first one is a concatenation layer which is used to combine the f_{sa}^r feature and f_{sa}^d feature. The second one is a 1×1 convolutional layer which is used to produce an $M^{2\times h\times w}$ modal mask. The last two are a softmax layer and a matrix multiplication layer. The former is used to generate a normalized modal mask and the latter is used for element-wise multiplication. For brevity and clarity, the layers are not illustrated in Figure 2.

To be more specific regarding the structure of the modality-aware module, after the concatenation layer, we obtain RGB-D fusion feature representation $f_f \in \mathbb{R}^{(2c\times h\times w)}$, and then feed it to the 1×1 convolutional layer to produce the mask M. Then M is fed into the softmax layer to produce a normalized modal mask $M' \in \mathbb{R}^{2\times h\times w}$. Let $M^{rgb} \in \mathbb{R}^{1\times h\times w}$ and $M^{depth} \in \mathbb{R}^{1\times h\times w}$ denote the modal masks on RGB and depth respectively. When the modal masks are generated, we calculate the predictions based on RGB and depth using the Hadamard product as follows:

$$\begin{aligned} P^{rgb} &= Conv(f_{at}^r) \odot M^{rgb} \\ P^{depth} &= Conv(f_{at}^d) \odot M^{depth} \end{aligned} \tag{3}$$

where $Conv(.)$ denotes a 1×1 convolutional layer, and $Conv(f_{at}^r) \in \mathbb{R}^{C\times h\times w}$, $Conv(f_{at}^d) \in \mathbb{R}^{C\times h\times w}$. The elements in $P(i,j)^{rgb}$ and $P(i,j)^{depth}$ imply how confidently we can rely on RGB and depth respectively to predict the pixel (i,j) in the input.

Finally, we generate the final prediction result as follows:

$$P^f = P^{rgb} + P^{depth} \tag{4}$$

3.3. Attention and Depth-estimation Modules

To further improve the segmentation performance, we propose attention and depth estimation modules to obtain long-range dependencies and more plausible feature representations. In order to enlarge the context relationship of the above-obtained f_{sa}^r and f_{sa}^d features, inspired by [33], we introduce a self-attention module to improve the obtained feature modeling ability.

For the sake of simplicity, let the size of the image feature obtained from the RGB modality scale-aware layer be $f_{sa}^r \in \mathbb{R}^{c\times h\times w}$, where c denotes the feature channel, $h = \frac{H}{8}$, $w = \frac{W}{8}$. Take the RGB modality as an example for explanation. We copy the f_{sa}^r and reshape it into three feature spaces, $\Theta(f) \in \mathbb{R}^{c\times N}$, $\theta(f) \in \mathbb{R}^{c\times N}$, and $\vartheta(f) \in \mathbb{R}^{c\times N}$, respectively, where $N = h \times w$. Then we calculate self-attention using the above two reshaped features, as follows:

$$at = softmax(\Theta(f)^T \cdot \theta(f)) \tag{5}$$

each item $at(i,j)$ in the module at is the dot-product similarity, which indicates the effect of the model at

the ith position to the jth position. To make it more implementation-friendly, we normalize the attention module before the softmax operation. Then we obtain the scale-aware attention feature representation as follows:

$$f_{at}^{r} = f_{sa}^{r} + \beta(at \cdot \vartheta(f)). \tag{6}$$

For the depth modality, f_{at}^{d} is similarly defined as f_{at}^{r}, where β is a learnable parameter, and is initialized to 0 during training inspired by [28]. The scheme mentioned above makes our model rely on non-attention features in the initial stages of training. For the depth modality, we utilize the same operation as the depth image feature representation.

For the depth estimation module, we adopt the structure of Monodepth2 [34], which is a successful depth estimation model. For simplicity of training, we adopt Depth Loss and Gradient Loss.

$$L_{depth} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\log^2(d_i) - \log^2(d_i^{Gt})} \tag{7}$$

$$L_{grad} = \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla(d) - \nabla(d^{Gt})\right\|_1 \tag{8}$$

where n is the number of pixels in the input image, d_i and d_i^{Gt} denote the predicted depth value and the corresponding ground truth depth value, respectively. In the experiments, the main purpose of the task is to obtain a per-pixel semantic segmentation label, and the depth estimation module is to encourage the RGB network to extract a more effective feature representation. We use the pre-trained Monodepth2 model to initialize our depth estimation module, and then use a small learning rate (1e-4) to fine-tune it in the final experiments.

4. Experiments

In this section, we perform extensive experiments on two publicly available datasets, NYU-Depth v2 and SUN RGB-D to evaluate our method. All of our implementations are made using the popular PyTorch framework.

4.1. Datasets

- NYU-Depth V2 is one of the most popular RGB-D indoor scene datasets, consisting of 1449 finely labeled RGB and depth image pairs. The entire dataset is divided into two parts, of which 795 are for training and 654 are for testing.
- SUN RGB-D is a large-scale RGB-D dataset recently used for indoor scene understanding. It contains 10335 pairs of RGB and depth images captured by four kinds of commercial depth sensors. Of these finely labeled image pairs, 5285 pairs are used for training and the remaining 5050 pairs are used for testing.

4.2. Metrics

Following recent methods [10, 35], performance in our experiments is quantitatively measured by pixel accuracy (Acc), mean intersection over union (mIoU), mean pixel accuracy of different categories (mAcc) and frequency weighted IoU (f.w. IoU), which are widely used in indoor semantic segmentation. To be concrete, let n_{ij} be the number of pixels which are misclassified as class j when the ground truth is category i. t_i is the number of pixels which belong to the ith category, where $t_i = \sum_j n_{ij}$, and the total number of pixels in the dataset is t. The above four metrics are defined as follows:

- pixel accuracy $= \sum_i \frac{n_{ii}}{t}$
- mean intersection over union $= \frac{1}{C}\sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$
- mean pixel accuracy $= \frac{1}{C}\sum_i \frac{n_{ii}}{t_i}$
- frequency weighted IoU $= \frac{1}{t}\sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

4.3. Training Protocol

In the following, we will provide details of the experimental implementation.

Learning rate policy The training procedure consists of two stages. In the first stage, we adopt the Adam optimizer to train two independent networks of RGB and depth modalities respectively for semantic segmentation, excluding the scale-aware and modality-aware modules. For each modality network, we adopt "poly" learning rate policy, where the current learning rate is calculated by multiplying the initial learning rate with $(1 - \frac{iter}{max\_iter})^{power}$, $power = 0.9$, the initial learning rate is set to 0.01. We use ResNet50 and ResNet101 as our backbone network and combine atrous spatial pyramid pooling as our feature encoder to extract multi-scale features. Each of the backbones is initialized by the model pre-trained on ImageNet, and the other layers are initialized by random weights. In the second stage, we add the scale-aware module and the modality-aware module and then fine-tune our RGB-D model on the synchronized RGB and depth training data. Each

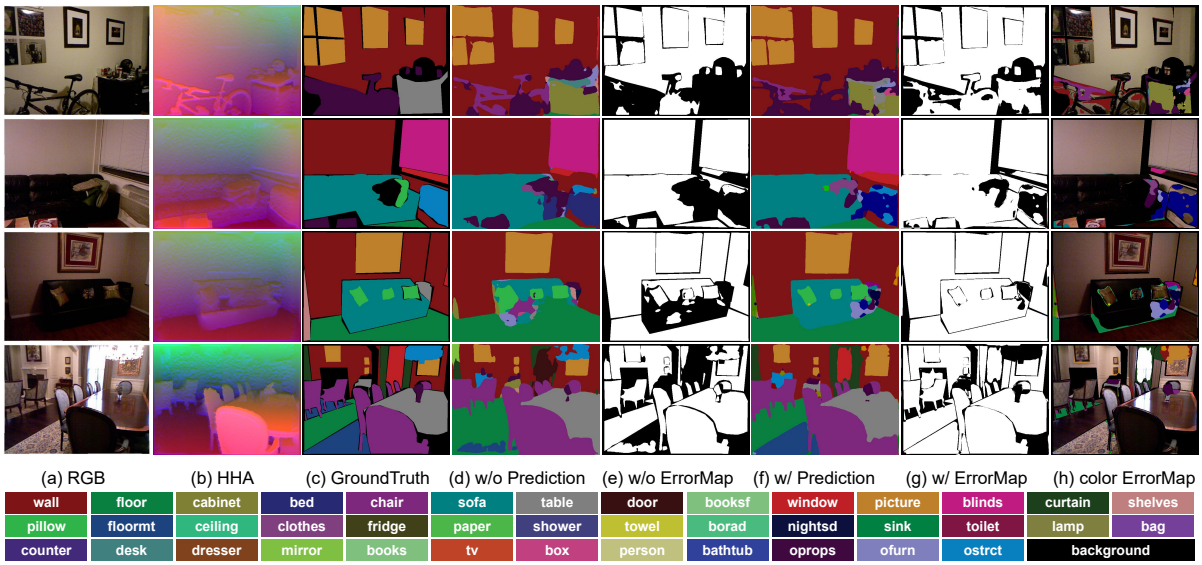


Figure 5: The visualization of results on the NYU-Depth v2 dataset. The comparison results of (d) and (f) demonstrate that our SAMD module is effective for indoor semantic segmentation. For the detailed analysis, please refer to Section 4.4. Best viewed in color.

modality network is initialized by the trained models obtained from the first stage. During the training, we discard the classification layer in the already-trained network in every single modality, and then combine them together via the added scale-aware and modality-aware modules. In the second stage, we set the initial learning rate to 0.001.

Data preprocessing and data augmentation In the experiment, the depth modality image is encoded to three-channel HHA (horizontal disparity, height above ground, and angle with gravity) image as the approach [6]. In both training stages, our two separate modality networks and our RGB-D model are trained on the cropped images of size 417×417. To avoid over-fitting, common data augmentations such as random brightness jittering, random left-right flipping, and random scaling in the range of [0.5, 2.0] to the input training samples are used.

Loss In the experiments, the overall loss is as follows:

$$L = L_{seg} + \lambda_1 \cdot L_{aux} + \lambda_2 \cdot L_{dep} \tag{9}$$

where λ_1 and λ_2 are the balancing weights for the semantic segmentation and depth estimation. To enhance the feature representation extracted from the backbone, we adopt an auxiliary loss after the 4th blocks (as used in [13]) to supervise the training process. In the experiments, λ_1 is set to 0.5, and λ_2 is set to 0.1. L_{dep} is composed by L_{Depth} and L_{grad}.

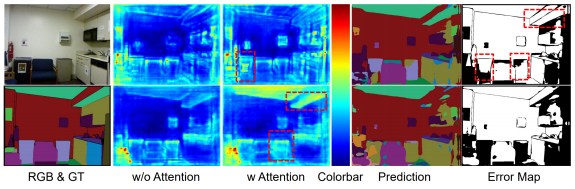


Figure 6: Illustration of the self-attention module. We observe that, with this module, the extracted feature representation is better (as shown in the red dashed bounding boxes). From left to the right and top to down, they are RGB and ground truth images; the RGB feature and HHA feature without attention module; RGB and depth feature with attention module; prediction results with and without attention module; error maps with and without attention module.

4.4. Ablation Studies and Discussion

In order to demonstrate that our SAMD model does not depend on any particular feature encoder architecture, we embed scale-aware module, attention module and modality-aware module into two standard fully convolutional backbone networks, ResNet50 and ResNet101. We provide the quantitative results on the NYU-Depth v2 dataset of these two backbone networks in Table 3. Through the results, we find that using our SAMD module significantly improves the performance of semantic segmentation throughout the different backbones. In the experiments, we use ResNet50 and ResNet101 as alternatives for our backbone network, and the default choice is ResNet101 if not explicitly specified.

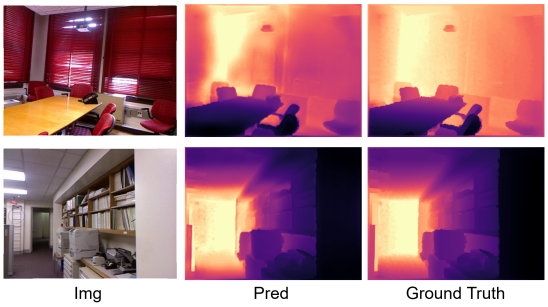
In order to show that our scale-aware module does not

Table 1: Category-wise IoU results on the NYU-Depth v2 dataset. The Baseline and SAMD rows show the results of our baseline and SAMD model respectively. The class of background is ignored during performance evaluation. The top two results are shown in red and blue respectively. Cheng‡ pre-trains their model on SUN RGB-D dataset, and then fine-tunes it on NYU-Depth v2 dataset.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floormat	clothes	ceiling
Long [7]	69.9	79.4	50.3	66.0	47.5	53.2	32.8	22.1	39.0	36.1	50.5	54.2	45.8	11.9	8.6	32.5	31.0	37.5	22.4	13.6	18.3	59.1
Gupta [6]	68.0	81.3	44.9	65.0	47.9	47.9	29.9	20.3	32.6	18.1	40.3	51.3	42.0	11.3	3.5	29.1	34.8	34.4	16.4	28.0	4.7	60.5
Deng [36]	65.6	79.2	51.9	66.7	41.0	55.7	36.5	20.3	33.2	32.6	44.6	53.6	49.1	10.8	9.1	47.6	27.6	42.5	30.2	32.7	12.6	56.7
He [37]	72.7	85.7	55.4	73.6	58.5	60.1	42.7	30.2	42.1	41.9	52.9	59.7	46.7	13.5	9.4	40.7	44.1	42.0	34.5	35.6	22.2	55.9
Cheng‡ [10]	78.5	87.1	56.6	70.1	65.2	63.9	46.9	35.9	47.1	48.9	54.3	66.3	51.7	20.6	13.7	49.8	43.2	50.4	48.5	32.2	24.7	62.0
Wang [35]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Daniel [38]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Gu [39]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Zhou [40]	80.1	88.3	61.7	72.8	63.9	65.4	48.0	46.5	48.3	44.4	61.4	69.9	59.5	27.2	16.8	59.3	50.6	50.9	51.3	38.6	25.1	79.5
Lin [41]	80.5	87.6	63.0	72.3	63.9	68.7	51.1	37.6	52.1	44.7	60.0	69.2	63.1	30.5	15.6	60.3	49.3	47.3	58.7	42.6	30.4	70.0
Cao [42]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Baseline	77.4	86.5	59.4	76.6	63.1	65.4	39.8	38.1	51.2	37.5	59.5	67.5	60.8	17.0	13.6	46.2	51.6	44.5	55.1	29.9	16.5	73.4
SAMD	82.3	89.7	62.3	73.0	64.8	67.7	51.0	46.8	51.8	46.9	65.5	71.3	62.2	23.4	19.7	60.1	48.8	49.7	51.7	42.0	26.6	81.2

	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathtub	bag	ostuct	ofurn	oprops	Acc	mIoU	mAcc	f.w. IoU
Long [7]	27.3	27.0	41.9	15.9	26.1	14.1	6.5	12.9	57.6	30.1	61.3	44.8	32.1	39.2	4.8	15.2	7.7	30.0	65.4	34.0	46.1	49.5
Gupta [6]	6.4	14.5	31.0	14.3	16.3	4.2	2.1	14.2	0.2	27.2	55.1	37.5	34.8	38.2	0.2	7.1	6.1	23.1	60.3	28.6	–	47.0
Deng [36]	8.9	21.6	19.2	28.0	28.6	22.9	1.6	1.0	9.6	30.6	48.4	41.8	28.1	27.6	0	9.8	7.6	24.5	63.8	31.5	–	48.5
He [37]	29.8	41.7	52.5	21.1	34.4	15.5	7.8	29.2	60.7	42.2	62.7	47.4	38.6	28.5	7.3	18.8	15.1	31.4	70.1	40.1	53.8	55.7
Cheng‡ [10]	34.2	45.3	53.4	27.7	42.6	23.9	11.2	58.8	53.2	54.1	80.4	59.2	45.5	52.6	15.9	12.7	16.4	29.3	71.9	45.9	60.7	59.3
Wang [35]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	43.9	53.5	–
Daniel [38]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.6	–	–
Gu [39]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	50.3	–	–
Zhou [40]	33.5	56.0	60.8	31.7	47.7	25.3	14.8	83.7	77.6	40.2	83.8	67.3	48.2	66.2	11.0	30.6	21.2	39.2	76.6	51.2	63.8	
Lin [41]	37.8	56.2	67.1	32.5	44.2	39.1	12.5	52.6	82.6	47.1	68.2	63.8	45.2	61.4	21.5	34.7	18.3	44.8	77.0	51.2	64.0	–
Cao [42]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	76.4	51.3	63.5	63.0
Baseline	24.9	45.8	53.2	23.2	39.8	27.1	5.1	73.5	64.9	38.4	86.2	67.5	43.3	57.0	5.1	28.0	19.9	38.6	73.4	46.7	61.7	61.2
SAMD	35.0	58.4	66.6	35.3	48.4	23.0	12.1	77.5	87.9	44.7	81.2	64.9	54.3	57.8	8.6	32.7	22.9	44.0	74.4	52.3	67.2	61.9

Figure 7: Performance Analysis. Depth estimation on NYUDepthv2 dataset.



depend on feature types, we utilize two methods, atrous spatial pyramid pooling (ASPP) and pyramid pooling (PSP) to extract multi-scale feature representations. In both experiments, we keep all settings exactly the same and extract four kinds of scale features in each modal network. We find that the use of ASPP (52.3) to extract multi-scale features is slightly better than PSP (52.1).

To demonstrate the effectiveness of the scale-aware module, we compare the results of using and not using this module. The qualitative analysis is shown in Figure 4. From the comparison result, with the scale-aware module, our model learns more proper scale feature representations for pixels. A pixel itself does not have enough contextual information for semantic segmentation, so it has to look around to check which class it belongs to. Whether it is from texture (RGB) or depth values (depth), the "board" object is very similar to surrounding pixels. The method of the forcible concatenation of the multi-scale feature would make some pixels confused when determining which category they belong to. Without the scale-aware module, the confidence map on the board region is low as shown in Figure 4. Also, the "wire" object (categorized into "otherprops") is too thin to be classified. With the scale-aware module, the model learns an appropriate feature representation. When comparing with the baseline model that does not have this module, the performance of our model is superior. To discover the importance of the self-attention module, we provide the comparison results with and without the module, as shown in Figure 6. From the results, we can find that, through the self-attention module, our model can model long-range dependencies.

The feature extracted from the scale-aware module, we can find that the feature extracted from the different atrous rate ai which is focused on the different region on the input images, as shown in Figure 3. From the

Table 2: Performance on the SUN RGB-D dataset. The SAMD row shows the results of our SAMD model. The class of background is ignored during performance evaluation.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror
Song [43]	36.4	45.8	15.4	23.3	19.9	11.6	19.3	6.0	7.9	12.8	3.6	5.2	2.2	7.0	1.7	4.4	5.4	3.1	5.6
Liu [44]	37.8	48.3	17.2	23.6	20.8	12.1	20.9	6.8	9.0	13.1	4.4	6.2	2.4	6.8	1.0	7.8	4.8	3.2	6.4
Ren [45]	43.2	78.6	26.2	42.5	33.2	40.6	34.3	33.2	43.6	23.1	57.2	31.8	42.3	12.1	18.4	59.1	31.4	49.5	24.8
Li [46]	74.9	82.3	47.3	62.1	67.7	55.5	57.8	45.6	52.8	43.1	56.7	39.4	48.6	37.3	9.6	63.4	35.0	45.8	44.5
Cheng [10]	91.9	94.7	61.6	82.2	87.5	62.8	68.3	47.9	68.0	48.4	69.1	49.4	51.3	35.0	24.0	68.7	60.5	66.5	57.6
Wang [35]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Zhou [40]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Cao [42]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SAMD	93.4	96.9	79.2	84.6	87.4	79.1	57.6	49.9	76.3	55.1	72.5	83.8	71.9	29.3	36.4	65.3	60.2	65.9	59.2

	floormat	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathhub	bag	mAcc
Song [43]	0.0	1.4	35.8	6.1	9.5	0.7	1.4	0.2	0.0	0.6	7.6	0.7	1.7	12.0	15.2	0.9	1.1	0.6	9.0
Liu [44]	0.0	1.6	49.2	8.7	10.1	0.6	1.4	0.2	0.0	0.8	8.6	0.8	1.8	14.9	16.8	1.2	1.1	1.3	10.1
Ren [45]	5.6	27.0	84.5	35.7	24.2	36.5	26.8	19.2	9.0	11.7	51.4	35.7	25.0	64.1	53.0	44.2	47.0	18.6	36.3
Li [46]	0.0	28.4	68.0	47.9	61.5	52.1	36.4	36.7	0.0	38.1	48.1	72.6	36.4	68.8	67.9	58.0	65.6	23.6	48.1
Cheng [10]	0.0	44.7	88.8	61.5	51.4	71.7	37.3	51.4	2.9	46.0	54.2	49.1	44.6	82.2	74.2	64.7	77.0	47.6	58.0
Wang [35]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	53.5
Zhou [40]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	60.5
Cao [42]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	58.5
SAMD	39.9	39.7	85.4	45.3	54.4	67.1	38.2	53.2	18.1	43.2	77.2	54.1	68.7	85.7	79.9	69.1	77.2	43.5	63.4

Table 3: Performance on different feature extractor encoder backbone network of our model.

Backbone	w/o SAMD	w/ SAMD
ResNet50	45.1	48.1
ResNet101	46.7	52.3

Table 4: Performance on different modality fusion methods of our model.

Methods	mIoU
Late fusion [7]	48.9
Gated fusion [10]	51.3
Modality-aware fusion	51.9

Table 5: Performance on the NYU-Depth v2 test dataset (4-class).

	Acc	mAcc
Courprie [5]	64.5	63.5
Hermans [47]	69.0	68.1
Stuckler [48]	70.6	66.8
Wang [49]	–	74.7
Eigen [50]	83.2	82.0
He [37]	83.6	82.5
SAMD	86.9	85.7

figure, we also find that the scale feature extracted from the scale-aware module, has different levels of attention on each modality. This phenomenon spurs us to design the next modality-aware module.

To demonstrate the effectiveness of the modality-aware module, we provide three results on the NYU-Depth v2 dataset with different modality fusion methods as shown in Table 4. In all three experiments, they all include the scale-aware and self-attention modules. All parameter settings in the experiments are the same except for the fusion method used. The late fusion approach follows the instruction in [7], which fuses RGB and depth networks by equal-weight score. [10] proposes a gated fusion way to fuse RGB and depth by regarding the varying contributions of the two complementary modalities. The last row is the performance of our modality-aware fusion, which achieves superior performance.

To demonstrate that the depth estimation module is workable and useful, we provide the depth estimation results of input images, as shown in Figure 7. From the results, we can find that the depth estimation can provide a plausible depth value for the input image.

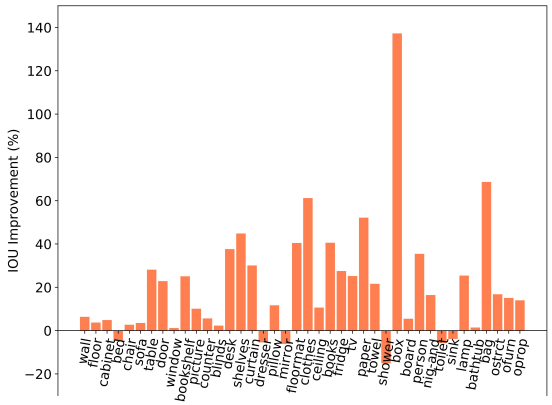
To have a better understanding of how the proposed SAMD model outperforms the baseline method, we provide the visualization results of the improvement of IoU for each semantic category in Figure 8. As can be seen from the statistics result in Figure 8, our SAMD is superior to the baseline in most classes.

In Table 6, we give the quantitative comparisons of with and without our SAMD components on the NYU-

Table 6: Ablation study of the proposed SAMD model on NYU-Depth v2 dataset. *S*, *A*, *M* and *D* denote scale-aware module, self-attention module, modality-aware module and depth estimation module, respectively.

	Methods	mIoU
a.	Baseline	46.7
b.	Baseline + *S*	47.8
c.	Baseline + *A*	48.6
d.	Baseline + *M*	48.4
e.	Baseline + *S* + *A*	49.8
f.	Baseline + *S* + *M*	49.9
g.	Baseline + *A* + *M*	49.5
h.	Baseline + *S* + *A* + *M*	51.9
i.	Baseline + *S* + *A* + *M* + *D*	52.3

Figure 8: Performance Analysis. Per-class IoU improvement of our SAMD model over baseline on NYU Depth-v2 test dataset.



Depth v2 dataset. From the comparison results ($b \sim i$), each component in the proposed SAMD module will benefit the performance of the indoor semantic segmentation. The qualitative results are illustrated in Figure 5, it gives the visualized comparisons with and without our SAMD module on the NYU-Depth v2 dataset. In Table 1, we give the results of the comparison between our model and state-of-the-art methods on the NYU-Depth v2 dataset. From the results, we can find that our model is better than state-of-the-art methods in many classes. We also test our model on the SUN RGBD dataset, and we obtain a state-of-the-art comparable result, 63.4% mean accuracy, more detail please refer to Table 2.

We compare SAMD to other state-of-the-art methods on the 4-class of the NYU-Depth v2 dataset, and the quantitative results are shown in Table 5. We also check the model size saved by PyTorch for both baseline and our SAMD to demonstrate the proposed module wouldn't increase the parameter size of the baseline too much. The size of baseline is 127.28M, and our SAMD is 136.19M.

5. Conclusion

In this paper, we propose SAMD to tackle the challenging problems for indoor semantic segmentation with RGB-D data. SAMD is composed of three main parts: (1) the scale-aware module which is designed for generating a spatial-sampled and scale-sampled feature representation, (2) the modality-aware module which can weigh the varying contributions of the two complementary modalities for better fusion, and (3) the self-attention module and depth estimation module, which can produce long-range dependencies for better modeling and push the RGB network to extract more plausible features. Theoretical analysis, qualitative and quantitative experimental results on NYU-Depth v2 and SUN RGB-D dataset demonstrate that SAMD can achieve significant performance gains for indoor semantic segmentation.

Acknowledgements

The authors thank the editor and all the reviewers for their very helpful comments to improve this paper. The authors also thank Professor Ou wu and Professor Xukun Shen for their suggestions and to polish this paper.

References

[1] Q. Xie, O. Remil, Y. Guo, M. Wang, M. Wei, J. Wang, Object detection and tracking under occlusion for object-level RGB-D video segmentation, IEEE Transactions on Multimedia 20 (3) (2017) 580–592.

[2] J. Huang, Z. Liu, Y. Wang, Joint scene classification and segmentation based on hidden Markov model, IEEE Transactions on Multimedia 7 (3) (2005) 538–550.

[3] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, S. Yan, Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval, IEEE Transactions on Multimedia 18 (6) (2016) 1175–1186.

[4] I. Ahn, C. Kim, Face and hair region labeling using semi-supervised spectral clustering-based multiple segmentations, IEEE Transactions on Multimedia 18 (7) (2016) 1414–1421.

[5] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, arXiv preprint arXiv:1301.3572.

[6] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: ECCV, 2014.

[7] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, PAMI.

[9] X. Q. X. W. J. J. Hengshuang Zhao, Jianping Shi, Pyramid scene parsing network, in: CVPR, 2017.

[10] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang, Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation, in: CVPR, 2017.

[11] X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, K. Yang, Gated fully fusion for semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11418–11425.

[12] X. Li, H. He, X. Li, D. Li, G. Cheng, J. Shi, L. Weng, Y. Tong, Z. Lin, Pointflow: Flowing semantics through points for aerial image segmentation, arXiv preprint arXiv:2103.06564.

[13] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587.

[14] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, arXiv preprint arXiv:2007.09183.

[15] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, H. Huang, Multi-scale context intertwining for semantic segmentation, in: ECCV, 2018.

[16] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, in: ICLR, 2015.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, arXiv preprint arXiv:1802.02611.

[18] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, J. Feng, Foveanet: Perspective-aware urban scene parsing, arXiv preprint arXiv:1708.02421.

[19] S. Kong, C. Fowlkes, Recurrent scene parsing with perspective understanding in the loop, arXiv preprint arXiv:1705.07238.

[20] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in: ICCV, 2021.

[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, arXiv preprint arXiv:1709.01507.

[22] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[23] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[24] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, D. Wierstra, Draw: A recurrent neural network for image generation, arXiv preprint arXiv:1502.04623.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML, 2015.

[26] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: CVPR, 2016.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.

[28] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, arXiv preprint arXiv:1805.08318.

[29] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: CVPR, 2018.

[30] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, L. Shao, Motion-attentive transition for zero-shot video object segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 13066–13073.

[31] L. Hoyer, D. Dai, Y. Chen, A. Koring, S. Saha, L. Van Gool, Three ways to improve semantic segmentation with self-supervised depth estimation, in: CVPR, 2021, pp. 11130–11140.

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

[33] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: CVPR, 2018.

[34] C. Godard, O. Mac Aodha, M. Firman, G. J. Brostow, Digging into self-supervised monocular depth estimation, in: ICCV, 2019, pp. 3828–3838.

[35] W. Wang, U. Neumann, Depth-aware CNN for RGB-D segmentation, in: ECCV, 2018.

[36] Z. Deng, S. Todorovic, L. Jan Latecki, Semantic segmentation of RGBD images with mutex constraints, in: ICCV, 2015.

[37] Y. He, W.-C. Chiu, M. Keuper, M. Fritz, STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling, in: CVPR, 2017.

[38] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, H.-M. Gross, Efficient RGB-D semantic segmentation for indoor scene analysis, arXiv e-prints (2020) arXiv–2011.

[39] Z. Gu, L. Niu, H. Zhao, L. Zhang, Hard pixel mining for depth privileged semantic segmentation, IEEE Transactions on Multimedia.

[40] H. Zhou, L. Qi, Z. Wan, H. Huang, X. Yang, RGB-D co-attention network for semantic segmentation, in: ACCV, 2020.

[41] D. Lin, H. Huang, Zig-zag network for semantic segmentation of RGB-D images, IEEE Transactions on Pattern Analysis and Machine Intelligence 2019 42 (10) (2020) 2642–2655.

[42] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, Y. Li, Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation, in: ICCV, 2021, pp. 7088–7097.

[43] S. Song, S. P. Lichtenberg, J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in: CVPR, 2015.

[44] C. Liu, J. Yuen, A. Torralba, Sift flow: Dense correspondence across scenes and its applications, IEEE transactions on pattern analysis and machine intelligence 33 (5) (2010) 978–994.

[45] X. Ren, L. Bo, D. Fox, RGB-(D) scene labeling: Features and algorithms, in: CVPR, IEEE, 2012, pp. 2759–2766.

[46] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, L. Lin, LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling, in: ECCV, 2016.

[47] A. Hermans, G. Floros, B. Leibe, Dense 3d semantic mapping of indoor scenes from RGB-D images, in: ICRA, IEEE, 2014, pp. 2631–2638.

[48] J. Stückler, B. Waldvogel, H. Schulz, S. Behnke, Dense real-time mapping of object-class semantics from RGB-D video, Journal of Real-Time Image Processing 10 (4) (2015) 599–609.

[49] J. Wang, Z. Wang, D. Tao, S. See, G. Wang, Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks, in: ECCV, Springer, 2016, pp. 664–679.

[50] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: ICCV, 2015, pp. 2650–2658.