Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation

Xiaokang Chen[1][0000−0002−6188−5821], Kwan-Yee Lin[2][0000−0003−0175−6398], Jingbo Wang[3][0000−0001−9700−6262], Wayne Wu[2][0000−0002−1364−8151], Chen Qian[2][0000−0002−8761−5563], Hongsheng Li[3][0000−0002−2664−7975], and Gang Zeng[1][0000−0002−9575−4651]

[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]SenseTime Research [3]The Chinese University of Hong Kong

Abstract. Depth information has proven to be a useful cue in the semantic segmentation of RGB-D images for providing a geometric counterpart to the RGB representation. Most existing works simply assume that depth measurements are accurate and well-aligned with the RGB pixels and models the problem as a cross-modal feature fusion to obtain better feature representations to achieve more accurate segmentation. This, however, may not lead to satisfactory results as actual depth data are generally noisy, which might worsen the accuracy as the networks go deeper.

In this paper, we propose a unified and efficient Cross-modality Guided Encoder to not only effectively recalibrate RGB feature responses, but also to distill accurate depth information via multiple stages and aggregate the two recalibrated representations alternatively. The key of the proposed architecture is a novel Separation-and-Aggregation Gating operation that jointly filters and recalibrates both representations before cross-modality aggregation. Meanwhile, a Bi-direction Multi-step Propagation strategy is introduced, on the one hand, to help to propagate and fuse information between the two modalities, and on the other hand, to preserve their specificity along the long-term propagation process. Besides, our proposed encoder can be easily injected into the previous encoder-decoder structures to boost their performance on RGB-D semantic segmentation. Our model outperforms state-of-the-arts consistently on both in-door and out-door challenging datasets [1].

Keywords: RGB-D Semantic Segmentation, Cross-Modality Feature Propagation

1 Introduction

Semantic segmentation, which aims at assigning each pixel with different semantic labels, is a long-standing task. Besides exploiting various contextual infor-

[1] Code of this work is available at https://charlescxk.github.io/

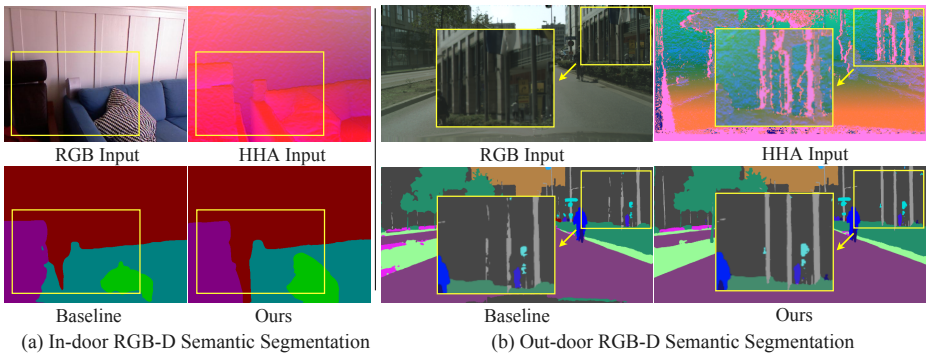

Fig. 1. (a)RGB-D baseline, which is designed with a habitual cross-modality fusion schema, results in inaccurate classification on the area that exists substantial variations between RGB and Depth modalities. (b) The depth measurements in out-door environments are noisy. Without proposed modules, the results will degrade dramatically

mation from the visual cues [24,11,14,12,5,41], depth data have recently been utilized as supplementary information to RGB data to achieve improved segmentation accuracy [25,31,42,4,15,22,6,19]. Depth data naturally complements RGB signals by providing the 3D geometry to 2D visual information, which is robust to illumination changes and helps better distinguishing various objects.

Although significant advances have been achieved in RGB semantic segmentation, directly feeding the complementary depth data into existing RGB semantic segmentation frameworks [24] or simply ensemble results of two modalities [6] might lead to inferior performance. The key challenges lie in two aspects. (1) *The substantial variations between RGB and Depth modalities.* RGB and depth data show different characteristics. How to effectively identify their differences and unify the two types of information into an efficient representation for semantic segmentation is still an open problem. (2) *The uncertainty of depth measurements.* Depth data provided with existing benchmarks are mainly captured by Time-of-Flight or structured light cameras, such as Kinect, AsusXtion and RealSense *etc.* The depth measurements are generally noisy due to different object materials and limited distance measurement range. The noise is more apparent for out-door scenes and results in undesirable segmentation, as shown in Fig 1.

Most existing RGB-D based methods mainly focus on tackling the first challenge. Standard practice is to use the depth data [2] as another input and adopt Fully Convolutional Network (FCN)-like architectures with feature fusion schemas, *e.g.,* convolution and modality-based affinity *etc.,* to fuse the features of two modalities [25,6,17,36]. The fused feature is then used to recalibrate the

[2] Raw depth map or its encoded representation–HHA map, which includes horizontal disparity, height above ground and norm angle. For more detail about HHA, please refer to [13].

subsequent RGB feature responses or predicted results. Although these methods provide plausible solutions to unify the two types of information, the assumption of the input depth data being accurate and well-aligned with RGB signals might not be true, making these methods sensitive to in-the-wild samples. Moreover, how to ensure that the network fully utilizes information from both modalities remains an open problem. Recently, some works [42,37] attempt to tackle the second challenge by diminishing the network's sensitivity to the quality of depth measurements. Instead of utilizing depth data as an extra input, they propose to distill the depth features via multi-task learning and regard depth data as extra supervision for training. Specifically, [37] introduces a two-stage framework, which first predicts several intermediate tasks including depth estimation and then uses the outputs of these intermediate tasks as the multi-modal input to final tasks. [42] proposes a pattern-affinitive propagation with jointly predicting depth, surface normal and semantic segmentation to capture correlative information between modalities. We argue that there exists an inherent inefficacy in such design, *i.e.* the interaction and correlation of RGB and depth information are only implicitly modeled. The complementarity of the two types of data for semantic segmentation was not well studied in this way.

Motivated by the above observations, we propose to tackle both two challenges in a simple yet effective framework by introducing a novel cross-modality guided encoder to FCN-like RGB-D semantic segmentation backbones. The key idea of the proposed framework is to leverage both channel-wise and spatial-wise correlation of the two modalities to firstly squeeze the exceptional feature responses of depth, which effectively suppresses feature responses from the low-quality depth measurements, and then use the suppressed depth representations to refine RGB features. In practice, we devise the steps bi-directionally due to the in-door RGB sources also contain noisy features. In contrast to depth data, the RGB noisy features are usually caused by similar appearance of different neighboring objects. We denote the above process as *depth-feature recalibration* and *RGB-feature recalibration*, respectively. We therefore introduce a new gate unit, namely the *Separation-and-Aggregation Gate (SA-Gate)*, to improve the quality of the multi-modality representation by encouraging the network to recalibrate and spotlight the modality-specific feature of each modality first, and then selectively aggregate the informative features from both modalities for the final segmentation. To effectively take advantage of the differences of features between the two modalities, we further introduce the *Bi-direction Multi-step Propagation (BMP)* that encourages the two streams to better preserve their specificity during the information interaction process in the encoder stage.

Our contributions can be summarized into three-fold:

- We propose a novel bi-directional cross-modality guided encoder for RGB-D semantic segmentation. With the proposed *SA-Gate* and *BMP* modules, we could effectively diminish the influence of noisy depth measurements, and also allow incorporating sufficiently complementary information to form discriminative representations for segmentation.

- Comprehensive evaluation on the NYUD V2 dataset shows significant improvements by our approach when integrated into state-of-the-art RGB semantic segmentation networks, which demonstrate the generalization of our encoder as a plug-and-play module.
- The proposed method achieves state-of-the-art performances on both in-door and challenging out-door semantic segmentation datasets.

2 Related Work

2.1 RGB-D Semantic Segmentation

With the development of depth sensors, recently there is a surge of interest in leveraging depth data as a geometry augmentation for RGB semantic segmentation task, dubbed as RGB-D semantic segmentation [25,31,20,23,42,3]. According to specific functionality of depth information suited in different architectures, current RGB-D based methods could be roughly divided into two categories.

Most of the works treat depth data as an additional input source to recalibrate the RGB feature responses either implicitly or explicitly. Long *et al.* [24] shows simply averaging final score maps of RGB and D modalities helps enforce the inter-object discrimination in the in-door setting. Li *et al.* [22] utilize the LSTM layers to selectively fuse the feature from the two modalities input. With a similar target, [6] proposes locality-sensitive deconvolution networks along with a gated fusion module. Several recent works [30,9,17] extend the RGB feature recalibration process from the final outputs of a dual-path network to different stages of the backbone, encouraging better recalibration with multi-level cross-modality feature fusion. To guide the recalibration with explicit cross-modality interaction modeling, some works [20,31,26,35] tailor general 2D operations to 2.5D behaviors with depth guidance. For example, [31] proposes depth-aware convolution and pooling operations to help recalibrating RGB feature responses in depth-consistent regions. [20] proposes a depth-aware gate module that adaptively selects the pooling field size in a CNN according to object scale. 3DGNN [26] introduces a 3D graph neural network to model accurate context with geometry cues provided by depth. Alternatively, some approaches regard the depth data as an extra supervised signal to recalibrate the RGB counterpart in a multi-task learning manner. For example, [42] proposes a pattern affinity propagation network to regularize and boost complementary tasks. [37] introduces a multi-modal distillation model to pass the valid messages from depth to RGB features.

Different from previous works that hold the ideal assumption of depth source's quality and mainly focus on in-door setting, we try to extend the task to the in-the-wild environment, *e.g.*, CityScapes dataset. The out-door setting is more challenging due to the inevitable noisy signals contained in the depth data. In this work, we try to recalibrate RGB feature responses from a filtered depth representation and vice versa, which effectively enhance the strength of representations for both modalities.

2.2 Attention Mechanism

Attention mechanisms have been widely utilized in kinds of computer vision tasks, serving as the tools to spotlight the most representative and informative regions of input signals [11,33,29,16,21,32]. For example, to improve the performance of the image/video classification task, SENet [16] introduces a self recalibrate gating mechanism by model importance among different channels of feature maps. Based on similar spirits, SKNet [21] designs a channel-wise attention module to select kernel sizes to adaptively adjust its receptive field size based on multiple scales of input information. [32] introduces a non-local operation which explores the similarity of each pair of points in space. For the segmentation task, a well-designed attention module could encourage the network to learn helpful context information effectively. For instance, DFN [39] introduces a channel attention block to select the more discriminative features from multi-level feature maps to get more accurate semantic information. DANet [11] proposes two types of attention modules to model the semantic inter-dependencies in spatial and channel dimensions respectively.

However, the main challenge of RGB-D semantic segmentation task is how to make full use of cross-modality data under the substantial variations and noisy signals between modalities. The proposed SA-Gate is the first to focus on the noisy features of cross-modalities by tailoring the attention mechanisms. The SA-Gate module is specialized for suppressing the exceptional noisy feature of depth data and recalibrate its counterpart RGB feature responses in a unified manner at first, and then fuses the cross-modality information with a softmax gating that is guided by the recalibrated features, achieving effective and efficient cross-modality feature aggregation.

3 Method

RGB-D semantic segmentation needs to aggregate features from both RGB and depth modalities. However, both modalities have inevitably noisy information. Specifically, depth measurements are inaccurate due to the characteristics of depth sensors and RGB features might generate confusing results due to the high appearance similarity between the objects. An effective cross-modality aggregation scheme should be able to identify their strengths from each feature as well as unify the most informative cross-modality features into an efficient representation. To this end, we put forward a novel cross-modality guided encoder. The overall framework of the proposed approach is depicted in Fig. 2 (a), which consists of a cross-modality guided encoder and a segmentation decoder. Given RGB-D data as inputs [3], our encoder recalibrates and fuses the complementary information from the two modalities via the SA-Gate unit, and then propagates the fused multi-modal features along with modality-specific features via the Bi-direction Multi-step Propagation (BMP) module. The information is then decoded by a segmentation decoder network to generate the segmentation map. We will detail each component in the remaining parts of this section.

[3] Note that we use HHA map to encode the depth measurements.

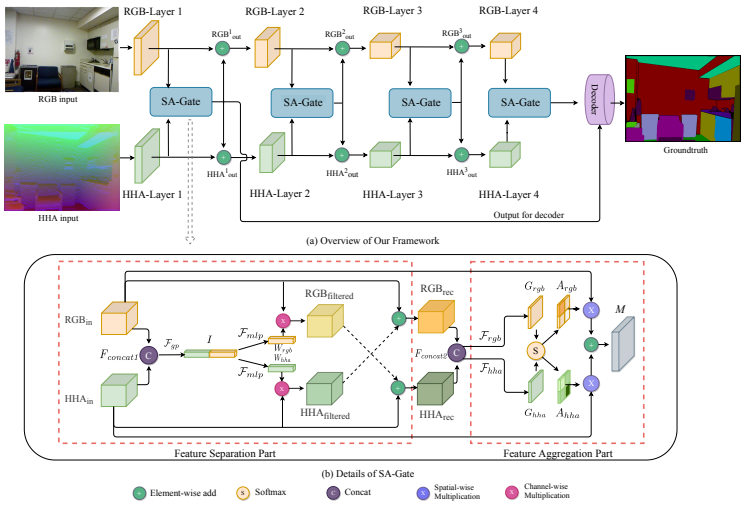

Fig. 2. (a)The overview of our network. We employ an encoder-decoder architecture. The input of the network is a pair of RGB-HHA images. During training, each pair of feature maps (*e.g.*, outputs of RGB-Layer1 and HHA-Layer1) are fused by a SA-Gate and propagated to the next stage of the encoder for further feature transformation. Fusion results of the first and the last SA-Gates would be propagated to the segmentation decoder (DeepLab V3+). (b) The architecture of the SA-Gate, which contains two parts, Feature Separation (FS) and Feature Aggregation (FA)

3.1 Bi-direction Guided Encoder

Separation-and-Aggregation (SA) Gate. To ensure informative feature propagation between modalities, the SA-Gate is designed with two operations. One is feature recalibration on each single modality, and the other is cross-modality feature aggregation. The operations are in terms of Feature Separation (FS) and Feature Aggregation (FA) parts, as illustrated in Fig 2 (b).

Feature Separation (FS). We take depth stream for example. Due to physical characteristics of depth sensors, noisy signals in depth modality frequently show up in regions close to object's boundaries or partial surfaces outside the scope of depth sensors, as shown in the second column of Fig. 3. Hence, the network is expected to first filter noisy signals surrounding these local regions to avoid misleading information propagation on the process of recalibrating complementary RGB modality and aggregating cross-modality features. In practice, we exploit high confident activations in RGB stream to filter out exceptional depth activations at the same level. To do so, global spatial information of both modalities should be embedded and squeezed to obtain a cross-modality attention vector first. We achieve this by a global average pooling along the channel-wise dimensions of two modalities, which is followed by concatenation and a MLP operation to obtain attention vector. Suppose we have two input feature maps denoted as $\mathrm{RGB_{in}} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ and $\mathrm{HHA_{in}} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, above operations could be formu-

lated as

$$I = \mathcal{F}_{gp}(\mathrm{RGB_{in}} \parallel \mathrm{HHA_{in}}), \tag{1}$$

where $\parallel$ denotes the concatenation of feature maps from two modalities, $\mathcal{F}_{gp}$ refers to global average pooling, $I = (I_1, \ldots, I_k, \ldots, I_{2C})$ is the cross-modality global descriptor for collecting expressive statistics for the whole inputs. Then, the cross-modality attention vector for the depth input is learned by

$$W_{hha} = \sigma(\mathcal{F}_{mlp}(I)), \quad W_{hha} \in \mathbb{R}^C, \tag{2}$$

where $\mathcal{F}_{mlp}$ denotes MLP network, σ denotes sigmoid function scaling the weight value into $(0, 1)$. By doing so, the network can take advantage of the most informative visual appearance and geometry features, and thus tends to effectively suppress the importance of noisy features in depth stream. Then, we could obtain a less noisy depth representation, namely Filtered HHA, through a channel-wise multiplication $\circledast$ between input depth feature maps and the cross-modality gate:

$$\mathrm{HHA_{filtered}} = \mathrm{HHA_{in}} \circledast \mathrm{W_{hha}}. \tag{3}$$

With a filtered depth representation counterpart, the RGB feature responses could be recalibrated with more accurate depth information. We devise the recalibration operation as the summation of the two modalities:

$$\mathrm{RGB_{rec}} = \mathrm{HHA_{filtered}} + \mathrm{RGB_{in}}, \tag{4}$$

where $\mathrm{RGB_{rec}}$ denotes recalibrated RGB feature maps. The general idea behind the formula is that, instead of directly using element-wise product to reweight RGB feature with regarding depth features as recalibrate coefficients, the proposed operation using summation could be viewed as some kind of offset to refine RGB feature responses at corresponding positions, as demonstrated in Table 2.

In practice, we implement *recalibration step* in a symmetric and bi-directional manner, such that low confident activations in RGB stream could also be suppressed in the same manner and filtered RGB information $\mathrm{RGB_{filtered}}$ could inversely recalibrate the depth feature responses to form a more robust depth representation $\mathrm{HHA_{rec}}$. We visualize feature maps of HHA before and after Feature Separation Part in Fig. 3. The RGB counterpart is shown in the supplementation.

Feature Aggregation (FA). RGB and D features are strongly complementary to each other. To make full use of their complementarity, we need to complementarily aggregate the cross-modality features at a certain position in space according to their characterization capabilities. To achieve this, we consider both characteristics of these two modalities and generate spatial-wise gates for both $\mathrm{RGB_{in}}$ and $\mathrm{HHA_{in}}$ to control information flow of each modality feature map with soft attention mechanism, which is visualized in Figure 2 (b) and marked by the second red frame. To make the gate more precise, we use recalibrated RGB and HHA feature maps from *FS* part, *i.e.*, $\mathrm{RGB_{rec}} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ and $\mathrm{HHA_{rec}} \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, to generate the gate. We first concatenate these two feature maps to combine their features at a certain position in space. Then we

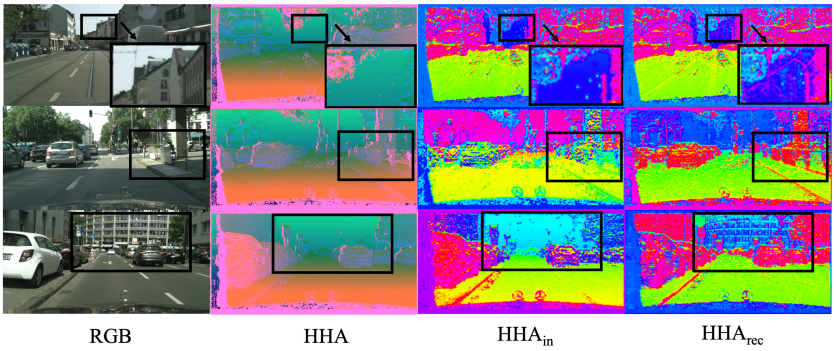

Fig. 3. Visualization of depth features before and after FSP on CityScapes validation set. We can observe that objects have more precise shapes after FSP and invalid partial surfaces are completed. More explanation is illustrated in the supplemental material

define two mapping functions to map high-dimensional feature to two different spatial-wise gates:

$$\mathcal{F}_{rgb} : F_{concat2} \rightarrow G_{rgb} \in \mathbb{R}^{1\times H\times W}, \tag{5}$$

$$\mathcal{F}_{hha} : F_{concat2} \rightarrow G_{hha} \in \mathbb{R}^{1\times H\times W}, \tag{6}$$

where $F_{concat2} \in \mathbb{R}^{2C\times H\times W}$ is the concatenated feature, G_{rgb} is the spatial-wise gate for RGB feature map, and G_{hha} is the spatial-wise gate for HHA feature map. In practice, we use a 1×1 convolution to implement this mapping function. A softmax function is applied on these two gates:

$$A_{rgb}^{(i,j)} = \frac{e^{G_{rgb}^{(i,j)}}}{e^{G_{rgb}^{(i,j)}} + e^{G_{hha}^{(i,j)}}},\ A_{hha}^{(i,j)} = \frac{e^{G_{hha}^{(i,j)}}}{e^{G_{rgb}^{(i,j)}} + e^{G_{hha}^{(i,j)}}} \tag{7}$$

where $A_{rgb}, A_{hha} \in R^{1\times H\times W}$ and $A_{rgb}^{(i,j)} + A_{hha}^{(i,j)} = 1$. $G_{rgb}^{(i,j)}$ is the weight assigned to each position in the RGB feature map and $G_{hha}^{(i,j)}$ is the weight assigned to each position in the HHA feature map. The final merged feature M can be obtained by weighting the RGB and HHA maps:

$$M_{i,j} = \mathrm{RGB}_{in}^{(i,j)} \cdot A_{rgb}^{(i,j)} + \mathrm{HHA}_{in}^{(i,j)} \cdot A_{hha}^{(i,j)}. \tag{8}$$

So far, we have added gated RGB and HHA feature maps to obtain the fused feature maps M. Since SA-Gate is injected into the encoder stage, we then average the fused features and the original input to obtain $\mathrm{RGB}_{\mathrm{out}}$ and $\mathrm{HHA}_{\mathrm{out}}$ respectively, which share similar spirits with residual learning.

Bi-directional Multi-step Propagation (BMP). By normalizing the sum of two weights at each position to 1, the numerical scale of the weighted feature will not significantly differ from the input RGB or HHA. Therefore, it has no negative influence on the learning of the encoder or the loading of the pre-trained

parameters. For each layer l, we use the output M^l generated by the l-th SA-Gate to refine the raw output of the l-th layer in the encoder: $\text{RGB}^l_{out} = (\text{RGB}^l_{in} + M^l)/2$, $\text{HHA}^l_{out} = (\text{HHA}^l_{in} + M^l)/2$. This is a bi-directional propagation process and the refined results will be propagated to the next layer in the encoder for more accurate and efficient encoding of the two modalities.

3.2 Segmentation Decoder

The decoder can adopt almost any design of decoder from SOTA RGB-based segmentation networks, since SA-Gate is a plug-and-play module and can make good use of complementary information of cross-modality on encoder stage. We show results of combining our encoder with different decoders in Table 6. We choose DeepLabV3+ [2] as our decoder for it achieves the best performance.

4 Experiments

We conduct comprehensive experiments on in-door NYU Depth V2 and out-door CityScapes datasets in terms of two metrics: mean Intersection-over-Union ($mIoU$) and pixel accuracy (pixel acc.). We also evaluate our model on SUN-RGBD dataset (Please refer to the supplemental material for more details).

4.1 Datasets

NYU Depth V2 [27] contains 1449 RGB-D images with 40-class labels, in which 795 images are used for training and the rest 654 images are for testing.
CityScapes [8] contains images from 27 cities. There are 2975 images for training, 500 for validation and 1525 for testing. Each image has a resolution of 2048×1024 and is fine-annotated with pixel-level labels of 19 semantic classes. **We do not use additional coarse annotations in our experiments**.

4.2 Implementation Details

We use PyTorch framework. For data augmentation, we use random horizontal flipping and scaling with scales [0.5,1.75]. When comparing with SOTA methods, we adopt flipping and multi-scale inference strategies as a test-time augmentation to boost the performance. More details are shown in the supplemental material.

Table 1. Comparison of efficiency on NYUDV2 test set. We use ResNet-50 as backbone and DeepLab V3+[2] as decoder. FLOPS are estimated for input of $3 \times 480 \times 480$

Methods	Params/M	FLOPs/G	mIoU(%)
RGB-D baseline	78.2	269.6	46.7
Ours	**63.4**	**204.9**	**50.4**

4.3 Efficiency Analysis

To verify whether the proposed cross-modality feature propagation helps and is efficient, we compare the final model with the RGB-D baseline. We average predictions of two parallel DeepLab V3+ as RGB-D baseline. As shown in Table 1, the proposed method achieves better performance with significantly less memory requirement and computational cost when compared with baseline. The results indicate that aimlessly adding parameters to a multi-modality network will not bring extra representational power to better recognize objects. In contrast, a well-design cross-modality mechanism, like proposed cross-modality feature propagation, helps to learn more powerful representations to improve performance more efficiently.

Table 2. Ablation study on *feature separation (FS)* part on NYU Depth V2 test set. No decoder is used here

Backbone	Concat	Self-global	Cross-global	Product	Proposed	mIoU(%)
Res50	✓					47.8
Res50		✓				47.5
Res50			✓			47.8
Res50				✓		47.5
Res50					✓	**48.6**

4.4 Ablation Study

We perform ablation studies on our design choices under same hyperparameters.
Feature Separation. We employ the FS operation before the feature aggregation in SA-Gate, to filter out noisy features for bi-directional recalibration step. To verify effectiveness of this operation, we ablate each design of FS in Table 2. Note that we ablate four different architectures and replace all FS parts in the network for comparison. 'Concat' represents we concatenate RGB_{in} and HHA_{in} feature maps and directly pass them to feature aggregation part. 'Self-global' represents we filter single modality features with its own global information. 'Cross-global' represents the filtered RGB is added to input RGB and vice versa. The filtering guidance comes from cross-modality global information. 'Product' means we multiply RGB_{in} by $\text{HHA}_{\text{filtered}}$ and vice versa. We see that from column 2 to 4, not using cross-modality information to filter noisy feature or refine features without explicit cross-modality recalibration lead to about 1% drop. On the other hand, the last two columns indicate the cross-modality guidance (E.q 4) is more appropriate and effective than cross-modality re-weighting when doing cross-modality recalibration. Overall, these results show that proposed FS operator effectively filters incorrect messages and recalibrates feature responses, achieving the best performance among all compared designs.
Feature Aggregation. We employ the SA-Gating mechanism to adaptively select the feature from the cross-modal data, according to their different characteristics at each spatial location. This gate can effectively control information flow

Table 3. Ablation study on *feature aggregation (FA)* part on NYU Depth V2 test set. No decoder is used here

Backbone	Addition	Conv	Proposed	mIoU(%)
Res50	✓			47.8
Res50		✓		48.0
Res50			✓	**48.6**

Table 4. Ablation study on encoder design on NYU Depth V2 test set. '*' means we average two outputs of RGB and HHA to get final output. No decoder is used here

Backbone	Block1	Block2	Block3	Block4	mIoU(%)
Res50*					45.9
Res50*	✓				47.8
Res50*		✓			47.5
Res50*			✓		46.8
Res50*				✓	44.3
Res50*	✓	✓			47.9
Res50*	✓	✓	✓		48.3
Res50*	✓	✓	✓	✓	48.0
Res50	✓	✓	✓	✓	**48.6**

of multimodal data. To evaluate the validity of the design, we perform ablation study on feature aggregation, as shown in Table 3. The experiment setting is kept the same as above. 'Addition' represents directly adding the recalibrated RGB and HHA feature maps. 'Conv' represents conducting convolution on the concatenated feature map. 'Proposed' represents the FA operator. We see that FA operator leads to the best result, since it considers the spatial-wise relationship between two modalities and can better explore the complementary information.

Design of Encoder. We verify and analyze the effectiveness of proposed BMP to our encoder, and how it functions with the SA-Gate. Toward this end, we conduct two ablation studies as shown in Table 4 & 5. We use ResNet-50 as our backbone here and directly upsampling the final score map by a factor of 16, without using a segmentation decoder. The first row in Table 4 & 5 is the baseline that averages score maps generated by two ResNet-50 (RGB & D).

For the first ablation, we gradually embed SA-Gate unit behind different layers of ResNet50. Note that we generate score maps for both two sides and average them as final segmentation result. This setting is different from those above, because last block of ResNet may not be equipped with a SA-Gate in this part, *i.e.*, no fused feature is generated from last block. From Table 4, we observe that if SA-Gate is embedded into a higher stage, it will lead to relatively worse performance. Besides, when stacking SA-Gate stage by stage, the additional gain continuously reduces. These two phenomena show that features of different modalities are more different in lower stage and an early fusing will achieve better performance. Table 5 shows results of second experiment. We observe that both SA-Gate and BMP can boost performance. Meanwhile, they complement each other and performs better in the presence of the other component. Moreover,

Table 5. Ablation study for BMP and SA-Gate. No decoder is used here

Method	mIoU(%)
Res50 (Average of Dual Path)	45.9
Res50 + SA-Gate	47.4 (1.5% ↑)
Res50 + BMP	47.8 (1.9% ↑)
Res50 + BMP + SA-Gate	**48.6** (2.7% ↑)

Table 6. The plug-and-play property evaluation of the proposed model on NYU Depth V2 test set. **Method** indicates different decoders, SA-Gate indicates the proposed fusion module. **RGB**: RGB image as inputs; **RGB-D**: the simple method which only average final score maps of RGB path and HHA path. Note that we reproduce these methods using official open-source code and all experiments use the same setting as our method

Method	RGB(%*mIoU*)	RGB-D(%*mIoU*)	RGB-D w SA-Gate(%*mIoU*)
DeepLab V3 [1]	44.7	46.5	**49.1** (2.6 ↑)
PSPNet [43]	43.1	46.2	**48.2** (2.0 ↑)
DenseASPP [38]	42.3	45.7	**47.8** (2.1 ↑)
OCNet [40]	44.5	47.6	**49.1** (1.5 ↑)
DeepLab V3+ [2]	44.3	46.7	**50.4** (3.7 ↑)
DANet [11]	43.0	45.5	**48.6** (3.1 ↑)
FastFCN [34]	45.4	47.6	**50.1** (2.5 ↑)

when associating Table 5 & 2, we see that SA-Gate helps BMP better propagate valid information than other gate mechanisms. It demonstrates effectiveness and importance of a more accurate representation to the feature propagation.

The Plug-and-Play Property of Proposed Encoder. We conduct ablation study to validate the flexibility and effectiveness of our method for different types of decoders. Following recent RGB-based semantic segmentation algorithms, we splice their decoders with our model to form modified RGB-D versions (*i.e.,* RGB-D w SA-Gate), as shown in Table 6. We see that in the column 2 and 4, our method consistently helps achieving significant improvements against original RGB versions. Besides, comparing with naive RGB-D modifications, our method also boosts the performance at least 1.5% *mIoU*. Especially, with the decoders in Deeplab V3+ [2], our method achieves 3.7% *mIoU* improvements. The results verify both the flexibility and effectiveness of our method for various decoders.

4.5 Visualization of SA-Gate

We visualize first SA-Gate in our model to see what it has learned, as shown in Fig 4. *Note that the black region in GT represents ignored pixels when calculating IoU. We reproduce RDFNet-101 [25] in PyTorch with 48.7% mIoU on NYU Depth V2, which is close to the result in the original paper (49.1%).* Red represents a higher weight assigned to RGB and blue represents a higher weight assigned to HHA. From column 4, we can see that RGB has a stronger response at boundary and HHA responds well in glare and dark areas. The phenomenon

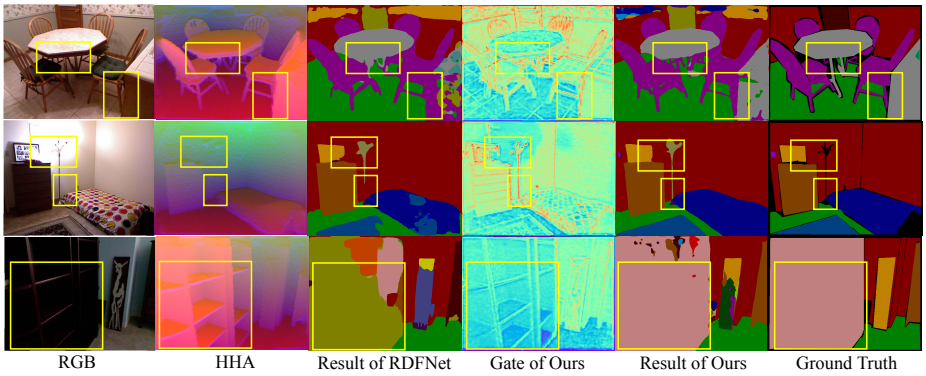

Fig. 4. Visualization of feature selection through SA-Gate on NYUD V2 test set. For each row, we show (1) RGB, (2) HHA, (3) results of RDFNet-101, (4) visualization of SA-Gate, (5) results of ours, (6) GT. Red represents a higher weight assigned to RGB and blue represents a higher weight assigned to HHA. Best viewed in color

Table 7. State-of-the-art comparison experiments on NYU Depth V2 test set

Method	mIoU(%)	Pixel Acc.(%)
3DGNN [26]	43.1	-
Kong *et al.* [20]	44.5	72.1
CFN [23]	47.7	-
RDF-101 [25]	49.1	75.6
PADNet [37]	50.2	75.2
ACNet [17]	48.3	-
PAP [42]	50.4	76.2
Ours	**52.4**	**77.9**

is reasonable since RGB feature has more details in high contrast areas and HHA feature is not affected by lighting conditions. From row 1, details inside yellow boxes are lost in HHA while obvious in RGB. Our method successfully identifies chair legs and distinguishes table that looks similar to chair. In row 2, glare blurs the border of the photo frame. Since our model focuses more on HHA in this area, it predicts the photo frame more completely than RDFNet. Besides, our model captures more details than RDFNet on clothes stand. In row 3, cabinet in dark red is hard to recognize in RGB but with identifiable features in HHA. Improper fusion of RGB and HHA leads to erroneous semantics for this area (column 3). While our model pays more attention to HHA in this area to achieve more precise results.

4.6 Comparing with State-of-the-arts

NYU Depth V2. Results are shown in Table 7. Our model achieves leading performance. On the consideration of a fair comparison to [42,17,37] that utilize ResNet-50 as backbone, we also use same backbone and achieve 51.3% $mIoU$,

Table 8. Cityscapes test set accuracies. '*' means RGB-D based methods

Method	roa.	sid.	bui.	wal.	fen.	pol.	lig.	sig.	veg.	ter.	sky	per.	rid.	car	tru.	bus	tra.	mot.	bic.	mIoU
CCNet [18]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	81.4
BFP [10]	98.7	87.0	93.5	59.8	63.4	68.9	76.8	80.9	93.7	72.8	95.5	87.0	72.1	96.0	77.6	89.0	86.9	69.2	77.6	81.4
DANet [11]	98.6	86.1	93.5	56.1	63.3	69.7	77.3	81.3	93.9	72.9	95.7	87.3	72.9	96.2	76.8	89.4	86.5	72.2	78.2	81.5
ACFNet [41]	98.7	87.1	93.9	60.2	63.9	71.1	78.6	81.5	94.0	72.9	95.9	88.1	74.1	96.5	76.6	89.3	81.5	72.1	79.2	81.8
LDFNet* [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.3
Shu Kong *et al.** [20]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	78.2
PADNet * [37]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	80.3
Choi *et al.** [7]	98.8	88.0	93.9	60.5	63.3	71.3	78.1	81.3	94.0	72.9	96.1	87.9	74.5	96.5	77.0	88.0	85.9	72.7	79.0	82.1
RGB baseline (Deeplab V3+ [2])	98.7	87.1	93.9	61.0	63.8	71.5	78.6	82.6	93.9	72.6	95.9	88.3	74.8	96.5	68.9	86.1	86.4	73.6	79.1	81.8
RGB-D baseline*	98.7	86.7	93.7	57.8	61.8	70.0	77.3	81.8	93.9	72.2	95.9	87.9	74.1	96.3	70.7	87.9	80.3	72.2	78.6	80.9
Ours*	98.7	87.3	93.9	63.8	62.7	70.8	77.9	82.2	93.9	72.8	95.9	88.2	75.2	96.5	80.4	91.6	89.0	73.2	78.9	**82.8**

which is still better than these methods. Specifically, [25,17] try to use channel-wise attention or vanilla convolution to extract complementary feature, which are more implicit than our model in selecting valid feature from complementary information. Besides, we can see that utilizing depth data as extra supervision (such as [42,37]) could make network more robust than general RGB-D methods that take both RGB and depth as input sources [25,6,26]. However, our results demonstrate that once the input RGB-D information could be effectively recalibrated and aggregated, higher performance could be obtained.

CityScapes. We achieve 81.7% *mIoU* on validation set and 82.8% *mIoU* on test set, which are both leading performances. Table 8 shows results on test set. We observe that due to serious noise of depth measurements in this dataset, most of previous RGB-D based methods even worse than RGB-based methods. However, our method effectively distills depth feature and extracts valid information in it and boosts the performance. Note that [7] is a contemporary work and we outperform them by 0.7%. We exclude the results of GSCNN [28] for fair comparison, since it uses a stronger backbone WideResNet instead of ResNet-101. However, we still outperform GSCNN by 0.9% mIoU on the validation set and achieve the same performance as it on test set.

5 Conclusion

In this work, we propose a cross-modality guided encoder along with SA-Gate and BMP modules to address two key challenges in RGB-D semantic segmentation, *i.e.,* the effective unified representation for different modalities and the robustness to low-quality depth source. Meanwhile, our proposed encoder can act as a plug-and-play module, which can be easily injected to current state-of-the-art RGB semantic segmentation frameworks to boost their performances.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (2017YFB1002601, 2016QY02D0304), National Natural Science Foundation of China (61375022, 61403005, 61632003), Beijing Advanced Innovation Center for Intelligent Robots and Systems (2018IRS11), and PEK-SenseTime Joint Laboratory of Machine Vision.

References

1. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
3. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: CVPR (2020)
4. Chen, Y., Mensink, T., Gavves, E.: 3d neighborhood convolution: Learning depth-aware features for rgb-d and rgb semantic segmentation. In: 3DV. IEEE (2019)
5. Cheng, B., Chen, L.C., Wei, Y., Zhu, Y., Huang, Z., Xiong, J., Huang, T.S., Hwu, W.M., Shi, H.: Spgnet: Semantic prediction guidance for scene parsing. In: ICCV (2019)
6. Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: CVPR (2017)
7. Choi, S., Kim, J.T., Choo, J.: Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
9. Deng, L., Yang, M., Li, T., He, Y., Wang, C.: Rfbnet: Deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. arXiv preprint arXiv:1907.00135 (2019)
10. Ding, H., Jiang, X., Liu, A., Thalmann, N.M., Wang, G.: Boundary-aware feature propagation for scene segmentation. In: ICCV (2019)
11. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019)
12. Fu, J., Liu, J., Wang, Y., Li, Y., Bao, Y., Tang, J., Lu, H.: Adaptive context network for scene parsing. In: ICCV (2019)
13. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: ECCV (2014)
14. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: CVPR (2019)
15. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In: ICCV (2017)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
17. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. arXiv preprint arXiv:1905.10089 (2019)
18. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
19. Hung, S.W., Lo, S.Y., Hang, H.M.: Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In: ICIP. IEEE (2019)
20. Kong, S., Fowlkes, C.C.: Recurrent scene parsing with perspective understanding in the loop. In: CVPR (2018)
21. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR (2019)

22. Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., Lin, L.: Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In: ECCV (2016)
23. Lin, D., Chen, G., Cohen-Or, D., Heng, P.A., Huang, H.: Cascaded feature network for semantic segmentation of rgb-d images. In: ICCV (2017)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
25. Park, S.J., Hong, K.S., Lee, S.: Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: ICCV (2017)
26. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: ICCV (2017)
27. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
28. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation (2019)
29. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
30. Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: ECCV (2016)
31. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: ECCV (2018)
32. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
33. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV (2018)
34. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv:1903.11816 (2019)
35. Xing, Y., Wang, J., Chen, X., Zeng, G.: 2.5 d convolution for rgb-d semantic segmentation. In: ICIP. IEEE (2019)
36. Xing, Y., Wang, J., Chen, X., Zeng, G.: Coupling two-stream rgb-d semantic segmentation network by idempotent mappings. In: ICIP. IEEE (2019)
37. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: CVPR (2018)
38. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: CVPR (2018)
39. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: CVPR (2018)
40. Yuan, Y., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)
41. Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E.: Acfnet: Attentional class feature network for semantic segmentation. In: ICCV (2019)
42. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: CVPR (2019)
43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)