A hybrid multi-modal visual data cross fusion network for indoor and outdoor segmentation

Sijie Hu*,†, Fabien Bonardi*, Samia bouchafa* and Désiré Sidibé*
*Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry-Courcouronnes, France
† Email: sijie.hu@universite-paris-saclay.fr

Abstract—Multi-modal scene parsing is a prevalent topic in robotics and autonomous driving since the knowledge of different modalities can complement each other. Recently, the success of self-attention-based methods has demonstrated the effectiveness of capturing long-range dependencies. However, the tremendous cost dramatically limits the application of this idea in multi-modal fusion. To alleviate this problem, this paper designs a multi-modal cross-fusion block (AC) and its elegant variant (EAC) based on an additive attention mechanism to capture global awareness among different modalities efficiently. Moreover, a simple yet efficient transformer-based trans-context block (TC) is also presented to connect the contextual information. Based on the above components, we propose light HCFNet, which can explore long-range dependencies of multi-modal information while keeping local details. Finally, we conduct comprehensive experiments and analyses on both indoor (NYUv2-13, -40) and outdoor (Cityscapes-11) datasets. Experiment results show that the proposed HCFNet achieved 66.9% and 51.5% mIoU on NYUv2-13 and -40 classes settings, which outperform current start-of-the-art multi-model methods. Our model also shows a competitive mIoU of 80.6% on the Cityscapes-11 dataset. The code will be available at https://github.com/Superjie13/HCFNet.

I. INTRODUCTION

As a fundamental task, semantic segmentation has received a broad range of attention in the computer vision community and industry. Depth information as an auxiliary provides shape and geometry cues of the surroundings that complement the RGB data, thus introducing depth information to improve the model’s performance has become a trend in robotics and autonomous driving. To this end, a series of networks with RGB-D as input appeared. These methods directly concatenate RGB and Depth images [1]–[3] or treat them in two branches [4]–[7].

Recently, self-attention-based transformer architecture has attracted attention in the computer vision community due to its flexibility in long-range modeling dependencies and its remarkable success in natural language processing (NLP). Therefore, well-designed transformer block (TB) or their variants are introduced to replace the region-wise convolution structure [8], [9], and their results have shown that a global view brought by self-attention helps draw a better perception structure [8], [10] or using a sliding window [9], [11]; others borrowed the idea from CNN-like architecture and introduce pyramid structure [12], [13]. Although various solutions for building long-range dependencies of individual RGB image reasoning emerged, the exploration of fusing RGB and Depth data for scene parsing is very limited. Existing multi-modal methods mainly deploy TB in a hybrid structure, i.e., mixing CNN and transformer, to ingest the advancement of both convolution and TB. The mainstream combinations of CNN and TB operations are (1) cascade: convolution operations are used to process high-resolution data and then followed by TBs to process low-resolution data. (2) parallel: CNNs are used as the backbone network for feature extraction, while TBs are independent modules to address the data fusion and exchange between different modalities. For example, [14] introduces the TB into the last two layers of the encoder in U-Net [15], reducing the calculation amount. [16] treat TB as an independent part and stack multiple self-attention modules to incorporate the global attention of the 3D scene. Compared with parallel structure, cascade structure fails to capture a larger context at the shallow level, and all existing methods rudely use the TB, so they inevitably bear the burden of the TB, i.e., the tremendous calculation.

We argue that the global attention founded by TB is the crucial reason for the success of Transformer. Recently, [17] proposed an efficient Transformer variant based on additive attention to achieve global attention modeling in linear complexity. Inspired by this, we propose a well-designed additive-attention-based cross-fusion block (AC) to incorporate depth information into RGB and form long-range dependencies between depth and RGB features. Besides, we present EAC block, an efficient variant of AC, which efficiently builds global contexts while maintaining fine-grained shape details. On the other hand, we offer a simple yet efficient trans-context module (TC) to enrich contextual information and capture a global context from fused features. Based on the above modules, we design a hybrid cross fusion network (HCFNet), as shown in Figure 1. With all the ideas, our method benefits from building global awareness while significantly reducing computational consumption. The formed global awareness crosses RGB and depth, bringing integrated information from different modalities. We report the experimental results on two commonly used datasets, namely NYUv2(-13, -40) [18] and Cityscapes-11 [19], to verify the effectiveness of the proposed method in both indoor and outdoor scenarios. The main contributions of this paper are summarized as follows:
We propose an efficient hybrid RGB-D data fusion network called HCFNet for semantic segmentation.

We propose a light data fusion block named additive attention cross fusion block (AC), and its variant (EAC), to form long-range dependencies cross depth and RGB features. Moreover, we offer a simple yet efficient trans-context module (TC) based on TB to build a global view of fused features.

We experimentally validate the proposed HCFNet on indoor and outdoor datasets, including NYUv2(-13, -40) and Cityscapes-11. Results show that our method achieves 66.9% mIoU on NYUv2-13, 51.5% mIoU on NYUv2-40, and 80.6% mIoU on Cityscapes-11 dataset, which is quite competitive compared with state-of-the-art RGB-D fusion methods.

II. RELATED WORK

A. Global attention and Transformer

Transformer was firstly proposed by [20] for NLP tasks. The core component is built upon multi-head self-attention, which can model the long-range dependencies within a sequence. A similar idea was introduced by [21] in computer vision to design a non-local block to build the global relationship between pixels. [22] proposed a full attention block based on a non-local block that computes global attention along both channel and spatial dimensions. As a pioneer of visual Transformer, [10] used pure Transformer structure to classify images and achieved promising results. [8] extended the work in [10] to semantic segmentation. Then, [9], [11] calculate self-attention in sub-windows to alleviate the resolution disaster, and apply Transformer structure to dense segmentation. [12], [13], [23] employed a pyramid or hierarchical Transformer structure to improve the computational efficiency of the model for segmentation. Moreover, [24] proposed a ‘transposed’ self-attention that computes global attention across feature channels so that the computational complexity is linear. Recently, [17] offered a variant of Transformer, in which additive attention replaces self-attention to establish global awareness. Compared with other global attention mechanisms, the calculation of additive attention is more efficient, so this article establishes a more general cross-modal fusion attention mechanism based on additive attention.

B. RGB-D semantic segmentation

With a more affordable depth sensor, semantic segmentation leveraged by the complementary geometric information of depth has drawn attention. However, the large noise in depth and the asymmetry between RGB and depth data make it challenging to integrate RGB and depth features effectively. In general, existing semantic segmentation structures include two stages: encoding and decoding. Concretely, input data are first encoded to form contextual feature embeddings then decoded to recover semantic information [25]. Some work [3], [26], [27] redesigned the convolution operation based on the characteristics of RGD-D data. [28] presented depth-aware operations to leverage depth similarity between pixels. [3] proposed a shape-aware convolutional layer. This convolutional layer is composed of two independent learnable components in the learning phase, and all the learnable parameters in the inference phase can be re-weight into a standard convolution operation. [26] introduced malleable 2.5D convolution to learn the receptive field along the depth axis. In contrast, most approaches are proposed to feed RGB and depth to two parallel branches [4], [5], [7], [29], [30]. For example, [31] employed two separate encoder-decoders to process RGB and depth, respectively, during which the manually designed gated fusion layer is used to fuse information from different streams. [32] used skip-connection to transmit the encoded multi-modal information to the decoder. [5]–[7] fuse the features...
at different stages of the encoding process. Recently, [30]
applied a shallow encoder and factorized convolutions to create
a lightweight model for real-time operations.

Unlike the above methods, we design a hybrid cross fusion
network that takes advantage of long-range dependencies in
the Transformer while maintaining the model’s efficiency.

III. METHOD

A. Overview of the method

An overview of our hybrid cross fusion network (HCFNet)
is presented in Figure 1. The structure is derived from a general
and classical multi-modal semantic segmentation paradigm,
i.e., two encoders for extracting features from RGB and Depth
and one decoder for reconstructing features from embeddings.
Similar to [5], [30], we use independent modules to achieve
data fusion of different modalities and pass features to the
decoder via skip-connections. The decoder is divided into
multiple stages. In each stage, feature maps are first treated
by a series of residual blocks [33] and then upsampled by
a factor of 2. The final output of the decoder is upsampled by
the factor of 4 to recover the original resolution. Our network
uses shallow encoders (i.e., ResNet-34 [33]) as the backbone
for feature extraction of both RGB and Depth streams to
reduce the footprint at runtime. In addition, we introduce
additive attention cross fusion blocks (AC) and EAC to fuse
valuable information efficiently during encoding and context-
block (TC) to enrich contextual features at the end of
the encoder.

B. AC block and its variant EAC

1) Additive attention: Additive attention was first intro-
duced in [17], which brings an effective global attention
mechanism to recalibrate the features within a sequence. A
basic form of additive attention is depicted in Figure 2. We
first summarize each token \((T_i, i \in [1 \ldots N])\) into an attention
scores by a linear transformation and a scale factor of \(\sqrt{d}\),
where \(d\) is the number of channels in a token. Then each
obtained attention score is normalized by a softmax operation
to get \(A_i^a\). The process can be formulated as:

\[
A_i^a = \frac{\exp(W_a^T T_i / \sqrt{d})}{\sum_{j=1}^N \exp(W_a^T T_j / \sqrt{d})},
\]

where \(N\) refers to the number of tokens and \(i \in [1 \ldots N]\),
\(W_a \in \mathbb{R}^{d \times d}\) are learnable weights of linear transformation. The
final global attention is obtained by weighted sum:

\[
A_i^g = f(T) = \sum_{i=1}^N A_i^a \cdot T_i.
\]

Note that additive attention has multiple heads as in the
standard self-attention.

2) AC block: As shown in Figure 3, the Additive attention
Cross fusion block (AC) adopts a symmetrical structure.
\(M_i \in \mathbb{R}^{d \times N}, i \in [1, 2]\) denote the inputs from two encoders.
Concretely, \(M_1\) and \(M_2\) are first processed by four linear
transformation (LT) units, respectively:

\[
M_i' = W_i^T M_i,
\]

where \(W_i \in \mathbb{R}^{d \times d}\) refers to learnable parameters in LT,
\(i \in [1, 2], j \in [1, 2, 3, 4]\). For the left part, \(M_1' \) is fed
into additive attention blocks to get a global attention score
\(A_i^{g1} \in \mathbb{R}^d\) and then element-wise multiplied by \(M_i'\) to
integrate attention score of \(M_1\) to the feature map of \(M_2\).
Then, in the same way, we build attention scores \(A_i^{g2}\), and
\(A_i^{g3}\) while only considering the feature map of \(M_1\). For the
right part, we use the same way to get \(A_i^{g4}\). Meantime, we also
introduce the information from \(M_1\) as an additional reference.
Note that we use the knowledge from another modality
to calibrate the long-range dependencies building process in
the current modality. This strategy makes it easier for AC block
to establish cross attention from one modality to another. This
process can be formalized as:

\[
A_i^{g5} = f(f(M_i^{g1}) \otimes M_i^{g2}_1) \otimes M_i^{g3},
\]

where \(f(T) = \sum_{i=1}^N A_i^a \cdot T_i\).
where \( i \in [1, 2] \), \( f \) denotes additive attention operation (see equation 2) and \( \otimes \) denotes element-wise multiplication operation.

At the same time, \( M_i^{1s} \) and \( M_2^{1s} \) are respectively transferred into two bypass branch modules to get \( M_i^1 \) and \( M_2^1 \). The bypass branch module is designed similar to the classical channel attention (CA) mechanism, which can be described as:

\[
\begin{align*}
M_i^c &= \text{Softmax}(M_i^1) \otimes M_i^{1s} \\
M_i^t &= W_i^cT M_i^c \\
M_i^r &= \text{AvePooling}(M_i^{1s}),
\end{align*}
\]

where \( i \in [1, 2] \), \( W_i^c \in \mathbb{R}^{d \times d} \) are the parameters of LT. \( \text{AvePooling} \) is average pooling operation along tokens.

Next, the generated \( A_i^{1s} \) and \( A_2^{1s} \) are respectively element-wise multiplied by \( M_i^{1s} \) and \( M_2^{1s} \), and then concatenated along channel axis to get \( M_i^{1c} \). Finally, after a linear transformation, \( M_i^{tc} \) is element-wise added with \( M_i^1 \) and \( M_2^1 \) to get the final output \( M_f^1 \):

\[
\begin{align*}
M_f^1 &= M_i^1 + M_2^1 + M_i^{1c} \\
M_i^{tc} &= W_i^{tc} T M_f^{tc},
\end{align*}
\]

where \( W_i^{tc} \in \mathbb{R}^{2d \times d} \) are parameters of LT and + refers to element-wise addition operation.

The idea behind this design is very intuitive. We use the global attention from \( M_1 \) to calibrate the features in \( M_2 \). Then the calibrated feature map of \( M_2 \) regenerates new global attention, which is further used to re-calibrate the features in \( M_1 \), and vice versa. Therefore, the block fully evaluates the interrelationship between different modalities to achieve a better efficient fusion. Note that AC block can also perform cross-feature fusion in sub-windows for more flexible analysis of local features.

3) **EAC block:** Efficient Additive attention Cross fusion block (EAC) is a variant of AC (section III-B2), designed to puzzle out the excessive consumption of building long-range dependencies under large resolution input. We consider that establishing global context information under full-resolution input will introduce redundancy, which leads to unnecessary calculations. In addition, fine-grained shape information is essential for establishing target contours. Accordingly, we decouple the process of establishing global context information and contour information. To do so, we design a shape extraction module in EAC block. An example for demonstrating the shape extraction module is shown in Figure 5. For each input, briefly, we compute the mean value of each local area by a sliding window with a certain stride which is equal to the size of the sliding window, to obtain the mean map. Then, the mean value in each local window is removed to obtain the shape map. Thanks to the parallel computing of Pytorch [34], this process can be implemented very efficiently.

EAC block is shown in Figure 4, we first build the mean map \( M^{im} \) and the shape map \( M^s \) of the input \( M_1 \) and \( M_2 \) through a shape extraction module. The extracted mean information \( M_1^{im} \) and \( M_2^{im} \) are input to the AC block to get the global context attention, and then followed by an upsampling operation to recover the resolution. The extracted shape information is processed by a pixel-wise convolution. Finally, mean information and shape information are integrated by element-wise addition. Table I shows the configuration of AC or EAC blocks at every encoding stage.

![Fig. 4. Structure of the proposed EAC block](image)

![Fig. 5. An example of Shape extraction in Figure 4](image)

![Fig. 6. Structure of the proposed TC block in Figure 1](image)
C. TC block

As shown in Figure 6, the Trans-Context block is composed of convolution and transformer blocks. Specifically, we first project the input channels through a $1 \times 1$ convolution, and then several TBs are applied to obtain complex context. Finally, we restore the number of input channels through another $1 \times 1$ convolution. The whole process is straightforward but very convenient. Note that the TB in our TC block can utilize existing well-designed methods, such as [10], [11]. We reimplement and employ TB of [17] in our TC block.

IV. EXPERIMENTS

A. Datasets and Implementation Details

1) NYUv2: NYUv2 is a popular dataset for indoor scene analysis. It contains 1449 indoor finely annotated RGB-D images, in which 795 are used for training and 654 for testing. All images are provided with a resolution of $640 \times 480$. We follow [3] using the train/test splits as provided by the dataset and report the results on the 13 and 40 classes [18] settings.

2) Cityscapes: The Cityscapes dataset is a large-scale database for urban street scene parsing. It contains 5000 finely annotated images captured from 50 cities with 19 semantic object categories, in which 2875 images are used for training, 500 images and 1525 images are used for validation and testing separately. All images are provided with a resolution of $2048 \times 1024$. We report results on the reduced 11 classes [4] setting.

3) Implementation Details: We implement our network based on Pytorch [34], and all experiments are run on a Nvidia RTX3090 GPU with 24GB memory. For the network, we take Resnet-34 initialized with the pre-trained weight on ImageNet [36] as the backbone of both encoders. We train our model for 500 epochs with a mini-batch size of 8 for the NYUv2 dataset and 300 epochs with a mini-batch size of 16 for the Cityscapes dataset. As for optimization, NYUv2 dataset is trained on SGD optimizer with a initial learning rate of 0.015 and Cityscapes dataset is trained on Adam optimizer with an initial learning of 0.0001. Following [30], we employ a one-cycle learning rate policy. Moreover, we set the number of TB in TC block ($N$) as 3. The image input size is set to $640 \times 480$ on the NYUv2 dataset and $768 \times 384$ on the Cityscapes dataset. If not otherwise noted, the inputs of all models are RGB and depth images. Note that before training on the Cityscapes dataset, we follow the official guide to generate a depth map from the original disparity data [19]. Random scaling, cropping, and flipping are applied for data augmentation to increase the number of training samples further. We evaluate our model based on mean intersection over union (mIoU). In addition, we still care for the frame per second (FPS) rate because of the computational burden.

B. Comparative results

1) Results on NYUv2: Table II compares the performance of our proposed methods with start-of-the-art methods. For a comprehensive comparison, we re-implement the prevalent multi-modal fusion methods based on their official repository and report the results on NYUv2-13 setting. Besides, we report our results on the commonly used NYUv2-40 setting. For the methods tested in the original paper, we use the reported results directly. We then follow [30] to modify our model by replacing BasicBlock with Non-Bottleneck-1D-Block (NBt1D) [35]. In our experiments, we also pay attention to FPS since they reflect the actual operating efficiency of the model. All FPS are executed at the input resolution of $480 \times 480$ on a laptop with Intel i7-9750 CPU and Nvidia RTX 2080 8G GPU. We noticed that the model based on NBt1D runs slower than the original model, which is inconsistent with the report in [30]. We consider this is because the $3 \times 3$ convolution is fully optimized in the computer environment. In table II we can see that our model outperforms current state-of-the-art methods on both NYUv2-13 and -40 classes settings while keeping a fast inference time.

In addition, we found that our method is capable of capturing overall contextual information while extracting valuable details. Please refer to the supplementary material for some qualitative results.

2) Results on Cityscapes: To exhibit the capabilities of our model in outdoor scenarios, we evaluate our model on the Cityscapes-11 dataset. Specifically, we resize the input image to a resolution of $768 \times 384$. All models are trained and evaluated at the same resolution. Note that all models are...
configured to the same training strategy, unless a different setting is provided in the original implementation. We observed that the Cityscapes dataset is very sensitive to the backbone, and a well pre-trained backbone can significantly improve the performance. As shown in Table III, our model yielded a very comparable result when using ResNet34 as backbone, and improved the segmentation results when using ResNet50 as backbone.

C. Ablation Analysis

To verify the functionality of the components of our model, we conduct an ablation study on the NYUv2-13 dataset. We use the network architecture in Figure 1 as the basic structure. For a fair comparison, network architecture and hyperparameters in different experiments are fixed. In ablations, we first evaluate the influence of different fusion methods, namely Add, ACM, SE, which have been used in recent start-of-the-art models. Specifically, Add [5], [29] simply adds the data of different modalities directly. ACM [7] and SE [30] calculate each modality’s attention through the Attention Complementary module (ACM) and Squeeze-and-Excitation module (SE) before fusion. Then we estimate the impact of the TC block on performance. Table IV summarizes the results of this ablation study on HCFNet.

From Experiment 1 to Experiment 3, we use different fusion methods to replace AC and EAC blocks of the original HCFNet while removing the TC block. As we can see, the proposed AC method is superior to other fusion methods by a large margin.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Fusion Method</th>
<th>Context</th>
<th>mIoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Add [5]</td>
<td>None</td>
<td>63.3</td>
</tr>
<tr>
<td>2</td>
<td>ACM [7]</td>
<td>None</td>
<td>64.1</td>
</tr>
<tr>
<td>3</td>
<td>SE [30]</td>
<td>None</td>
<td>63.9</td>
</tr>
<tr>
<td>4</td>
<td>AC* (Ours)</td>
<td>None</td>
<td>65.2</td>
</tr>
<tr>
<td>5</td>
<td>AC* (Ours)</td>
<td>TC</td>
<td>65.8</td>
</tr>
<tr>
<td>6</td>
<td>AC (Ours)</td>
<td>TC</td>
<td>65.1</td>
</tr>
</tbody>
</table>

* denotes we follow Table I to configure AC and EAC blocks.

In addition, after deploying the TC block, the performance of our model is significantly improved (+0.6%), which reveals that it is practical to further establish long-range dependencies in the fused features. In the final experiment, we replaced the EAC block in Table I with the AC block, which caused a decrease in model performance. This echoes our previous argument that the context contains much redundant information under large-resolution input, which will lead to unnecessary calculations and confusion. In contrast, our proposed AC block can efficiently establish long-range awareness, and with the help of the EAC block, redundant information can be eliminated without destroying local details while significantly reducing the amount of calculation. Finally, the proposed TC block can further coordinate information fusion and establish a more significant receptive field.

V. CONCLUSION

In this paper, we designed a novel multi-modal visual data fusion method, which can efficiently integrate data from different modalities. It also ensures that the model retains valuable local details after fusion while having a global receptive field. Precisely, we customized a multi-modal fusion block named AC block based on the additive attention mechanism, which assists form global awareness inter- and inner-modalities. Then, we proposed the EAC block, an efficient variant of the AC block, to efficiently build global attention and keep details under high-resolution input. On the other hand, based on the transformer block, we offered a simple yet effective context fusion block called trans-context (TC) block for further connecting the context output from the encoder. Together with the proposed well-designed components, we present HCFNet for semantic segmentation of indoor and outdoor scenarios. Finally, comprehensive experiments and ablation studies verified the effectiveness of our network and different components.

In the future, we will further optimize the structure of the model and exploit the possibilities of our model on different other modalities such as infrared or polarimetry.
REFERENCES


