UCTNet: Uncertainty-aware Cross-modal Transformer Network for Indoor RGB-D Semantic Segmentation

Xiaowen Ying and Mooi Choo Chuah

Lehigh University
xiy517@lehigh.edu, chuah@cse.lehigh.edu

Abstract. In this paper, we tackle the problem of RGB-D Semantic Segmentation. The key challenges in solving this problem lie in 1) how to extract features from depth sensor data and 2) how to effectively fuse the features extracted from the two modalities. For the first challenge, we found that the depth information obtained from the sensor is not always reliable (*e.g.* objects with reflective or dark surfaces typically have inaccurate or void sensor readings), and existing methods that extract depth features using ConvNets did not explicitly consider the reliability of depth value at different pixel locations. To tackle this challenge, we propose a novel mechanism, namely Uncertainty-Aware Self-Attention that explicitly controls the information flow from unreliable depth pixels to confident depth pixels during feature extraction. For the second challenge, we propose an effective and scalable fusion module based on Cross-Attention that can adaptively fuse and exchange information between the RGB encoder and depth encoder. Our proposed framework, namely UCTNet, is an encoder-decoder network that naturally incorporates these two key designs for robust and accurate RGB-D Segmentation. Experimental results show that UCTNet outperforms existing works and achieves state-of-the-art performances on two RGB-D Semantic Segmentation benchmarks.

1 Introduction

Semantic Segmentation is a task that aims to gain pixel-level understandings of the scene. Given an input RGB image, the goal of Semantic Segmentation is to classify each pixel into a set of predefined semantic categories. A single monocular RGB image can be seen as a 2D projection of a 3D scene. During the imaging procedure, the information in the depth dimension is inevitably lost. With the development of sensor technology, depth sensors are becoming widely accessible and can help recover the missing information in the depth dimension which is valuable for scene understanding.

In this paper, we focus on the task of depth-assisted (RGB-D) Semantic Segmentation. There are two major challenges in this task: 1) how to effectively extract features from the additional depth input (since feature extraction from

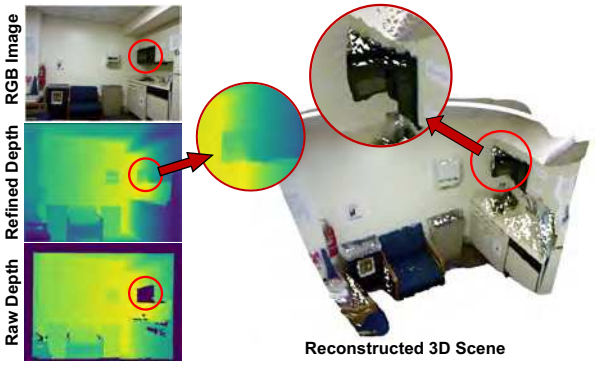

Fig. 1. An example RGB-D image and its reconstructed 3D scene. The raw depth map has no valid measurements on the surface of the microwave (highlighted in Red Circle) and such missing values are filled by certain algorithms (Refined Depth). These uncertain area looks reasonable in the refined depth map but leads to a largely distorted reconstruction and therefore should not be completely trusted.

RGB images has been extensively studied for decades), and 2) how to aggregate and fuse the features extracted from two input modalities.

For the first challenge, existing approaches typically consider depth maps as single-channel images and employ CNNs (Convolutional Neural Networks) to extract features from the depth map similar to the RGB input. However, such approaches omit an important characteristic of the depth sensor, *i.e.* not every depth value in the depth maps is reliable. Most of the depth sensors available nowadays, either based on *structured light*, *ToF*, or *Lidar*, rely on measuring the reflection of the light signals they sent out. Due to the physical constraint, it is difficult to correctly measure the depth values on some surfaces such as glasses or dark materials. Figure 1 illustrates an example where we can see the raw depth map does not have valid measurements on the microwave since it has a reflective surface. To avoid feeding the *raw depth map* with missing values to the Neural Networks, a common practice is to use the *refined depth* map as input instead, which is generated by filling the missing values using certain algorithms such as colorization [30]. However, as the filled values are estimated by algorithms, we should not completely trust them. We can see from Figure 1 that even though the refined depth map itself looks reasonable, the resulting reconstructed object in the 3D space is largely distorted.

To tackle this problem, our goal is to design a framework that explicitly considers the reliability of the input values during feature extraction. This goal may not be easily achieved using traditional CNNs since Convolutions are input-agnostic (always applies the same set of kernels during inference regardless of different inputs), and we found that the recently developed Vision Transformers (ViTs) are more suitable for achieving our goal. Instead of using Convolution operations to extract features, ViTs are built upon the Self-Attention (SA) oper-

ations which are input-specific [37]. A SA operation can be seen as propagating information on a fully connected *undirected graph* where nodes are the pixels and edges are their attentions. Based on this point-of-view, we proposed a novel mechanism called Uncertainty-Aware Self-Attention (UASA), by re-modeling the attention operation as a *directed graph* and explicitly controlling the information flow coming out from the uncertain nodes. In other words, UASA allows uncertain nodes to accept information from other confident nodes but limits the information it sends out. We replace all the SA operations in our depth encoder with the proposed UASA, and our experimental results demonstrate its effectiveness compared to the traditional SA operation for extracting features from the depth map.

The second challenge is how to fuse the information extracted from two input modalities. We review and analyze the pros and cons of existing fusion strategies in the literature and summarize four design principles. Following these principles, we design a new fusion module that can perform adaptive and asymmetric information exchange between two branches. Our fusion module is based on the Cross-Attention (CA) technique that aligns well with our ViT backbone and we proposed two modifications to make it scalable to high-resolution feature maps and easier to train. We demonstrate the effectiveness of our fusion module compared to baselines and other fusion strategies with our ablation studies in Section 4.3.

Our final framework, namely UCTNet, is an encoder-decoder network that incorporates our proposed two designs for RGB-D Semantic Segmentation. Since the transformer-based backbones are shown to be more powerful feature extractors compared to traditional CNN-based backbones [32], we also perform careful ablation studies to demonstrate the effectiveness of our contributions over the strong baselines. Finally, we evaluate our framework on two public benchmarks for RGB-D Semantic Segmentation and show that UCTNet significantly outperforms previous approaches on both benchmarks.

Our contributions can be summarized as follow:

- We introduce a novel Uncertainty-Aware Self-Attention mechanism to explicitly handle the feature extraction from inputs with uncertain values.
- We design an effective and scalable fusion module that can perform adaptive and asymmetric information exchange between two branches.
- Our proposed framework, namely UCTNet, achieves new state-of-the-art performance on two public benchmarks and outperforms all existing methods with significant improvements.

2 Related Work

Semantic Segmentation. Traditional Semantic Segmentation takes as input an RGB image and aims to predict every pixel in the image into a set of predefined categories. FCN [33] proposed one of the first deep learning-based semantic segmentation frameworks by replacing the fully-connected layer in a deep image

classification model with a convolution layer to support pixel-wise classification. Following works [36,1,17,3,49] explored to add different types of decoder networks and skip-connections to produce finer segmentation results. This architecture, which is also referred to as the *Encoder-decoder Network*, is the most popular architecture for semantic segmentation and is still being used in many state-of-the-art approaches. Following the encoder-decoder architecture, a line of works [31,56,19,23,13,24] explore to incorporate multi-scale analysis to the semantic segmentation network. Another line of works [5,6,7,8] explores the use of Dilated Convolutions to increase the receptive field while maintaining similar computational costs.

RGB-D Semantic Segmentation. A depth map provides complementary information to the corresponding RGB image that helps recover the information in the missing dimension. Earlier works [22,29] have shown that adding depth information can improve the segmentation results. As we discussed in the previous section, extracting features from depth maps and incorporating the features from two modalities are not trivial problems since the depth maps have different input distribution and characteristics compared to RGB images. To solve these problems, a line of works [52,54,47,9,11] try to design special "depth-aware" convolution operations to handle the depth information. These new operations can be seen as the augmented version of convolution and they technically have similar complexities compared to the original convolution; however, they usually run much slower in practice due to the lack of efficient and optimized implementation. Another line of works [22,29,26,53,10,38,46,16,41] simply employ a dual-encoder design, in which two separate encoders are used to extract features from the RGB image and depth map, respectively. Most of the state-of-the-art approaches follow this dual-encoder architecture as it allows different encoders to focus on extracting modality-specific features and typically yields better performance. However, none of the aforementioned approaches explicitly consider the uncertainty of the depth map — those modified convolutions only added "depth-aware" functionality to the convolution, and the dual-encoder network typically employs the same encoder structure for both branches and lets the network learn the modality-specific features implicitly. Our proposed framework follows the dual-encoder design but has a specifically designed uncertainty-aware encoder for the depth modality.

RGB-D Fusion. For all dual-encoder approaches that extract features from two modalities using separate encoders, a key problem is how to fuse and combine modality-specific features from two encoders. Within the scope of RGB-D Semantic Segmentation, early works [22] [29] adopt a naive fusion strategy by fusing the depth features to the RGB encoder using element-wise addition. Seichter et al. [41] perform fusion using channel-wise weighted addition where the weight is produced by a Squeeze-and-Excitation (SE) module. However, their fusion weights are generated from each input feature, respectively, and are not adaptive to both inputs. The fusion module in [10] uses attentive addition and produces the fusion weight by considering both input features. They additionally pass the combined feature back to both encoders to enhance not only the RGB

encoder but also the depth encoder. However, their fusion module only outputs one combined feature, meaning that both encoders receive the same fused feature regardless of the input modality they process. This problem inspires us to come up with the *Asymmetric* principle in Section 3.3.

Fusion techniques in other RGB-D-related tasks are typically not directly compatible with our framework. [28] fuses the features from RGB and depth encoders into a third encoder-decoder network which introduces high computational overheads. Fusion techniques in [35,44,40,4,18] all involve customization of the entire decoding stage and hence are not compatible with other existing semantic decoders. Our fusion module follows the modular design principle and is compatible with most of the well-designed encoders and decoders in the existing Semantic Segmentation literature.

Vision Transformers. Convolutional Neural Networks (CNNs) have been the most popular architecture in building the encoder-decoder architecture for semantic segmentation in the past decade. Recently, a novel architecture called Vision Transformers (ViTs) has attracted much interest in the Computer Vision community. The Transformer has proven to be a very powerful feature extractor in the Natural Language Processing (NLP) problems and was recently introduced to the Computer Vision tasks [14,45,55,12,21,48]. At this point, the major problem of ViTs is the high computational cost as the Self-Attention operation has quadratic complexity. Liu et al. [32] solved this problem by substituting the Self-Attention operation with their proposed Shifted Window Self-Attention (SWSA) which reduces the complexity to linear. While our proposed method is compatible with any ViTs, we choose to use the Swin Transformer [32] as the base architecture of our encoder network for it is one of the first ViTs that are both powerful and efficient.

3 Proposed Method

Due to the physical constraint, most depth sensors available nowadays cannot obtain valid readings on reflective or light-absorbent surfaces such as glasses and mirrors, and the prediction of a model can be severely affected if it happens to make decisions based on information in uncertain regions. Our goal is to design a mechanism that allows the model to *explicitly* consider the uncertainty during feature extractions. This goal (explicitly handling the uncertainty) may not be easily achieved using traditional CNNs since Convolutions are input-agnostic, meaning that the features are always extracted using the same set of kernels regardless of different inputs. To enable adaptive feature extraction based on the uncertainty, we develop our approach based on the recent popular ViT architectures since their core operation, namely Self-Attentions (SA), are *input-specific* [37]. The overall architecture of our framework is described in Section 3.1. Section 3.2 describes how we modify the Self-Attention to handle depth uncertainty. Our fusion module is described in Section 3.3.

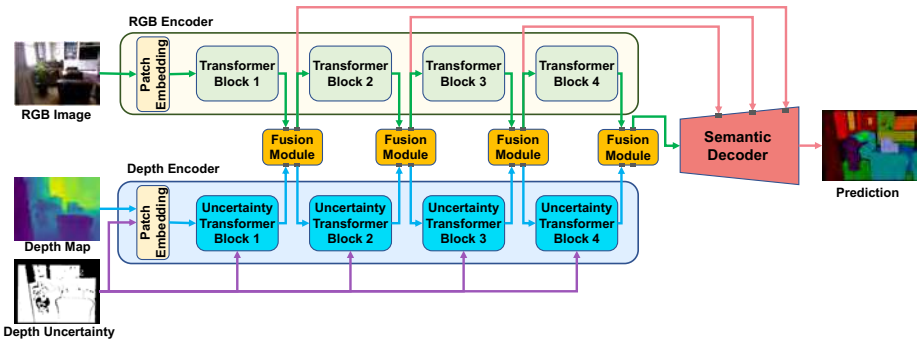

Fig. 2. The architecture of our proposed framework.

3.1 Overall Architecture

The framework of UCTNet, as illustrated in Figure 2, follows the dual-encoder design used in state-of-the-art approaches — it consists of two parallel encoders to extract modality-specific features from the image and depth modalities, respectively, followed by a semantic decoder to generate the final segmentation results. For the RGB encoder, we adapt the Swin-S [32] architecture which is a powerful and efficient ViT backbone. Given an input RGB image, the Swin-S backbone first splits the image into small patches (each patch is 4×4 in our case) and generates a patch embedding for each patch via a linear embedding layer. The patch features will go through four sequential transformer blocks that produces image features in $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ resolutions, respectively. The depth encoder adopts the same architecture as the RGB encoder, except that 1) we substitute all Self-Attention (SA) layers in its transformer blocks with our proposed Uncertainty-Aware Self-Attention (UASA) layer, and 2) we additionally feed the depth uncertainty map (describe in Section 3.2) to the Patch Embedding layer by concatenating it with the input depth map. More details of the UASA layer are described in Section 3.2.

At the output of each encoder stage, we use our fusion module to fuse and exchange information between the RGB encoder and the depth encoder. The fusion module takes inputs from both RGB and depth branches and returns the updated features back to their corresponding encoder for the next stage. More details of our fusion module are described in Section 3.3.

One may notice that the logical structure of the Swin Transformer is very similar to the ResNets [25], in which the Patch Embedding layer is similar to the "Stem" layers in ResNet and the four Transformer blocks are analogous to the four corresponding ResNet blocks. Therefore, it is possible to attach any existing semantic decoder compatible with a ResNet-like encoder to construct an encoder-decoder network for semantic segmentation. Similar to [32], we use the UperNet decoder [51] as our semantic decoder in this paper for its efficiency. More details of the architecture can be found in our supplementary materials.

3.2 Uncertainty-Aware Self-Attention

Self-Attention (SA). Self-attention is the core layer in a Vision Transformer Encoder. Unlike a Convolution layer that extracts features from a local kernel, self-attention allows information propagation between every single pair of input features. Given an input feature map $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$, the traditional Self-Attention operation is computed as:

$$SA(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V, \tag{1}$$

where $Q,K,V \in \mathbb{R}^{H\times W\times d}$ are the *query*, *key* and *value* produced by mapping each C-dimensional feature in $\mathcal{X}$ to d-dimensional embeddings via three different linear layers, respectively.

One can see that SA can be interpreted as a fully-connected graph $\mathcal{G}(\mathcal{X},\mathcal{E})$ where each feature in $\mathcal{X}$ is a node and the edge between two nodes are the attention weights between two features computed by QK^T in equation 1. After each SA operation, each node in the graph gets updated by gathering information flow from other nodes it connects to, and the edge values (attention weights) control how much information it should keep from different nodes. This graph is also an undirected graph since $\mathcal{E}_{i\to j} = \mathcal{E}_{j\to i}$, meaning that the attention weight from node i to node j is the same as the other way around.

Depth Uncertainty. Depth measurements from the sensors could be affected by the physical environment. In general, existing depth sensors typically have difficulties in measuring depth for surfaces that are highly reflective or have high light absorption. Traditional depth sensors such as Kinect simply return a void value if the depth cannot be accurately measured. In these cases, we consider its uncertainty map as a binary map $\mathcal{U} \in \{0,1\}^{H\times W}$ where zeros denote no sensor reading at this location and one denotes a valid sensor reading. Some newer sensors can provide multi-level confidence maps (e.g. three-level confidence map in Apple's Lidar Scanner). In such cases, we can normalize the confident map to $\mathcal{U} \in [0,1]$. In the following subsection, we formulate our Uncertainty-Aware Self-Attention for any generic type of uncertainty map $\mathcal{U} \in [0,1]$ consists of either binary, multi-level discrete, or continuous values. It is worth mentioning that the current two major benchmarks for RGB-D Semantic Segmentation are collected using traditional sensors such as Kinect which only provides binary uncertainty maps. Yet our experimental results show that the performance is improved significantly by incorporating such simple uncertainty into the framework.

Uncertainty-Aware Self-Attention. The proposed Uncertainty-Aware Self-Attention (UASA) operation extends the SA operation by considering a bidirected graph $\mathcal{G}(\mathcal{X},\mathcal{E})$ where $\mathcal{E}_{i\to j} \neq \mathcal{E}_{j\to i}$. In this way, we can explicitly control the information flow between a confident node and an uncertain node. Here, we introduce two variants of UASA: **Cut-off** and **Suppression**.

UASA (Cut-off). Given an uncertainty map $\mathcal{U} \in [0,1]^{H\times W}$, the cut-off variant of the UASA can be computed as:

$$UASA_{cut}(Q,K,V) = softmax(\frac{QK^\top}{\sqrt{d}} - M)V, \tag{2}$$

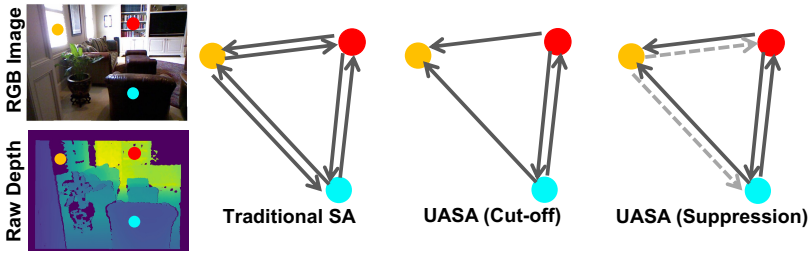

Fig. 3. Illustration of the Uncertainty-Aware Self-Attention among three selected pixels (🔴🟡🔵) in an example scene. A traditional SA always has $A_{i,j} = A_{j,i}$. UASA (Cut-off) simply discards the information flows from uncertain nodes (🟡) to confident nodes (🔴🔵), while UASA (Suppression) keeps these edges but reduce their attention weights.

where $M \in \mathbb{R}^{HW \times HW}$ and:

$$M_{i,j} = \begin{cases} 0, & \text{if } \mathcal{U}_j \geq \theta \\ \infty, & \text{if } \mathcal{U}_j < \theta \end{cases} \tag{3}$$

The $UASA_{cut}$ operation simply cuts off the outward information flow from all uncertain nodes in which their confidences are less than a threshold θ, such that their features will not be propagated to other nodes. However, they can still receive information from other confident nodes and use them to update their node features.

UASA (Suppression). The $UASA_{cut}$ operation may be too aggressive since the information from those uncertain nodes may still be useful. On one hand, as we mentioned in Section 1, their initial values can be filled by using an estimation algorithm. On the other hand, as the input goes through multiple transformer layers, those uncertain nodes already got updated multiple times by features from confident nodes, and hence their uncertainties are reduced. To this end we consider a softer variant of UASA:

$$UASA_{sup}(Q, K, V) = softmax(\frac{QK^\top \cdot \mathcal{S}}{\sqrt{d}})V, \tag{4}$$

where $\mathcal{S} \in \mathbb{R}^{HW \times HW}$ and:

$$\mathcal{S}_{i,j} = \frac{1}{\mathcal{T} + \mathcal{U}_j \cdot (1 - \mathcal{T})}, \tag{5}$$

where $\mathcal{T}$ is a hyper-parameter indicating the maximum temperature (corresponding to those nodes with zero confidences).

Instead of simply cutting off all the information from uncertain nodes, Eq. 4 and 5 suppress these uncertain features by dividing their attention weights with a temperature (calculated based on the node uncertainty) before applying the Softmax operation. Figure 3 provides a high-level illustration of the UASA using three selected pixels.

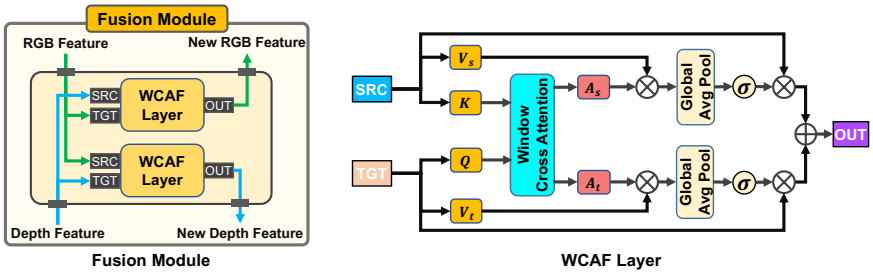

Fig. 4. The architecture of the proposed Window Cross Attentive Fusion (WCAF) layer.

UASA for Shift-Window Attention. The above formulations define our UASA based on the traditional Self-Attention operation, while the Swin-S backbone we employed for our encoder uses a modified version of SA named Shift-Window Self-Attention (SWSA). However, the above formulations also work for SWSA as we simply need to do the same window-partition operation on uncertainty map $\mathcal{U}$ along with the input image before we perform UASA on each window. We omit the detail of the window-partition operation in our formulation for simplicity and refer readers to its original paper [32].

3.3 Fusion Module

The design goal of our fusion module is to achieve feature fusion and information exchange between two encoding streams. We choose not to modify the decoder to allow our approach compatible with any existing decoder.

Based on the analysis of existing fusion modules in Section 2, we summarize the following design principles for our fusion module:

- **Attentive:** The features from different modalities should be combined in an attentive way instead of simple element-wise addition.
- **Adaptive:** The attention/weight to perform attentive fusion should be generated by adaptively considering both input modalities.
- **Bidirectional:** Instead of one-way passing the feature from one modality to another, we prefer to exchange the information between two modalities.
- **Asymmetric:** The combined features passing back to different encoders should be different, *i.e.* $F_{depth \to rgb} \neq F_{rgb \to depth}$, where F denotes the fusion function.

Apart from these design principles, we also prefer to design the fusion module using the Attention mechanism to align with our ViT backbone. The architecture of our Fusion Module is illustrated in Figure 4. The core layer in our fusion module, namely the Window Cross Attentive Fusion (WCAF) layer, takes a *source* feature and a *target* feature as inputs, and the goal is to fuse the information from *source* feature into the *target* feature. The *source* and *target*

features may come from either RGB or depth modalities depending on the fusion direction. The WCAF layer is based on the Cross-Attention mechanism with two key modifications: (1) the vanilla cross-attention has quadratic complexity which scales badly to high-resolution features, and we proposed Window Cross-Attention (inspired by [32]) that has linear complexity and can be used to fuse high-resolution features from the early stages of the encoder. (2) it is very difficult to learn a Cross-Attention layer that performs well to directly fuse dense features. We address this by combining cross-attention with channel-attention so that the cross-attention only needs to produce channel weighting which greatly reduces the training difficulties.

The design of WCAF layer met our first two design principles, *i.e. Attentive* and *Adaptive*. To further achieve the last two principles, we employ two independent WCAF layers in our fusion module (as shown in Figure 4) to approximate $F_{depth \to rgb}$ and $F_{rgb \to depth}$, respectively. These two WCAF layers have identical architectures except the input order is reversed. This allows us to enhance the features in both RGB and depth encoders while keeping them independent to extract their own modality-specific features.

4 Experiments

4.1 Implementation Details

Architecture. We implement and train our networks using the PyTorch [39] framework. We use the default configuration of Swin-S [32] for our encoders except we change the input channel in the first layer of the depth encoder to match its input modality. The input resolutions for both RGB and depth encoders are 640×480, and the output feature channels at four intermediate stages are $\{96, 192, 384, 768\}$.

Losses. The main training loss is a pixel-wise Cross-Entropy loss with *frequency class balancing* strategy [15] applied to the final decoder output. During training, we follow the common practices to attach an auxiliary FCN head to the stage 3 features that produce an auxiliary prediction at 1/16 resolution and apply the same Cross-Entropy loss to this output as an auxiliary loss. The final loss for optimization is a weighted sum of the main loss and the auxiliary loss, where the weights are set to 1.0 and 0.4, respectively.

Training. As in [32], we employ the AdamW [34] as our optimizer during training with a learning rate of 6e-6. During training, we initialize the network with weights pretrained on ADE20K datasets provided by [32] except for the first layer of the depth encoder which is randomly initialized (since the input channel is different). All the experiments in this paper are trained for 500 epochs with batches of size 2.

4.2 Datasets

We train and evaluate the performance of our networks on two public datasets for indoor RGBD Semantic Segmentation — NYUv2 [42] and SUN RGB-D [43].

NYUv2. NYUv2 dataset is comprised of RGB-D images taken from a variety of indoor scenes. The raw depth image is captured using a Microsoft Kinect sensor and the refined depth is generated using the colorization scheme proposed in [30]. It provides 1,449 densely labeled data which is split into a training set of 795 samples and a testing set of 654 samples. The annotations are provided in 13, 40, and 894 class settings but we adopt the most-common 40-class setting as in most of the existing works.

SUN RGB-D. SUN RGB-D is a large-scale benchmark for RGB-D scene understanding tasks. It not only consists of a significant amount of newly captured data but also combines samples from multiple existing datasets including NYUv2 [42], Berkeley B3DO [27] and SUN3D [50]. SUN RGB-D consists of 10,335 indoor RGB-D images which are split into a training set with 5,285 samples and a testing set of 5,050 samples. All images are densely annotated with 37-classes semantic labels.

4.3 Ablation Study

Baselines. Our framework employs a Vision Transformer backbone for this task while most of the existing works use CNN-based backbones. To reveal the real contribution of our proposed approach, we first design two baselines for our ablation study. The first baseline model (*Baseline-1* in Table 1) is a typical encoder-decoder network for semantic segmentation using *RGB* images as input (w/o depth image). The second baseline model (*Baseline-2*) uses the same architecture as Baseline-1 except that it takes *RGB-D* images as input by simply concatenating the RGB image with the depth map along the channel axis, resulting in a 4-channel input image. Note that these two baseline models do not involve depth encoders and hence no fusion module is needed. The result of *Baseline-1* demonstrates the power of ViT backbones for the Semantic Segmentation task since the performance it achieves using merely RGB image is already a very strong baseline. *Baseline-2* is a naive solution for RGB-D segmentation by simply concatenating RGB image and depth image as a 4-channel input and processing it with the same encoder-decoder architecture. The result shows that the segmentation performance benefits from this additional information but the improvement is minor.

Different Fusion Module. Once we start adding depth encoder to extract better depth-specific features, we will need a fusion module to exchange information between the two encoders. The *Ablation 1-4* in Table 1 compare our fusion module with three existing fusion modules used in recent state-of-the-arts. We can see that the result using our fusion module (*Ablation-4*) is better than the other three existing fusion modules, which demonstrates the effectiveness of our proposed fusion module. It is worth mentioning that the design of SA-Gate [10] meets the first three design principles yet its performance is not as good as expected. This may be caused that the design of SA-Gate being optimized for the HHA [20] depth encoding instead of a normal single-channel depth image.

Uncertainty-Aware Transformer Encoder. One of our core contributions in this paper is the Uncertainty-Aware Encoder that explicitly considers

Method	RGB Enc.	Depth Enc.	Fusion	mIoU (%)
Baseline-1 (RGB only)	TR-Enc.	N/A	N/A	52.2
Baseline-2 (RGB cat. D)	TR-Enc.	N/A	N/A	52.4
Ablation-1	TR-Enc.	TR-Enc.	Element-Add [22,29]	54.3
Ablation-2	TR-Enc.	TR-Enc.	SE-Add [41]	54.6
Ablation-2	TR-Enc.	TR-Enc.	SA-Gate [10]	53.8
Ablation-3	TR-Enc.	TR-Enc.	Our Fusion	55.3
Ablation-4	TR-Enc.	UATR-Enc.(Cut-off)	Our Fusion	56.8
Ablation-5 (Final Model)	TR-Enc.	UATR-Enc.(Suppression)	Our Fusion	**57.6**

Table 1. Ablation Study on NYUv2 dataset. **TR-Enc.**: Encoder consists of standard transformer blocks. **UATR-Enc.**: Encoder consists of Uncertainty-Aware transformer blocks.

Value of $\mathcal{T}$	5	10	**15**	20
mIoU(%)	57.2	57.2	**57.6**	56.8

Table 2. Sensitivity Analysis of the Temperature $\mathcal{T}$ in UASA.

the input uncertainties during feature extraction. As shown in Table 1 *Ablation 3-5*, replacing the standard encoder with our Uncertainty-Aware encoder for the depth branch leads to significant performance improvements. We can also see that the *Suppression* variant of our UASA performs better than the *Cut-off* variant, which supports our hypothesis that the features from those uncertain nodes are not completely unreliable and should not be completely dropped out. Therefore, we choose the *Suppression* variant of our UASA in our final model.

Choice of Temperature $\mathcal{T}$ in UASA. A key hyperparameter in our UASA (Suppression) is the temperature that controls the suppression strength for the outward information from uncertain nodes. Intuitively, if the temperature is too small, the information from uncertain nodes may interfere with and mess up the features in the confident nodes; if the temperature is too large, the information from uncertain nodes will be almost dropped out, which could lead to negative impacts as we discussed in previous ablation studies. Table 2 shows the sensitivity analysis of the temperature $\mathcal{T}$ in the UASA module. We can see that the performance improves by increasing $\mathcal{T}$ from 5 to 15 but starts to drop when $\mathcal{T}$ increases to 20. We also notice that the performance of $\mathcal{T} = 20$ is the same as the *Cut-off* version of UASA in Table 1, which implies that $\mathcal{T} = 20$ is a fairly strong suppression factor that almost cuts off most of the uncertain information.

4.4 Comparison with State-of-the-arts

We compare our proposed framework with existing state-of-the-art methods on two public benchmarks — Table 3 lists all the results on the NYUv2 dataset the SUN RGB-D benchmark. On the NYUv2 dataset, our approach outperforms all existing methods with significant improvement where we achieve 7.1% absolute

Method	Category	Architecture	NYUv2	SUN RGB-D
2.5D Conv [54]	Mod. Conv	1 × R101	48.5	48.2
SGNet [9]	Mod. Conv	1 × R101	49.0	47.1
ShapeConv [2]	Mod. Conv	1 × R50	47.3	46.3
ShapeConv [2]	Mod. Conv	1 × R101	50.2	47.6
FuseNet [22]	Dual Enc.	2 × VGG16	-	37.3
RedNet [29]	Dual Enc.	2 × R34	-	47.8
SSMA [46]	Dual Enc.	2 × R50	-	44.4
ACNet [26]	Trio Enc.	3 × R50	48.3	48.1
MMAF-Net [16]	Dual Enc.	2 × R152	44.8	47.0
Idempotent [53]	Dual Enc.	2 × R101	49.9	47.6
RDFNet [38]	Dual Enc.	2 × R152	50.1*	47.7*
ESANet [41]	Dual Enc.	2 × R50	50.5	48.3
SA-Gate [10]	Dual Enc.	2 × R101	52.4*	49.4*
UCTNet (Ours)	Dual Enc.	2 × Swin-S	**57.6**	**51.2**

Table 3. Quantitative results (MIoU) on the NYUv2 dataset and SUN RGB-D benchmark compared to state-of-the-art RGB-D Semantic Segmentation methods. Mod. Conv: Methods based on modified convolutions. Dual Enc.: Methods based on dual encoder architectures. *: Multi-scale testing.

improvement in terms of mIoU compared to the previous best method without using multi-scale testing.

The SUN RGB-D benchmark is a large-scale benchmark with much more training and testing samples compared to the NYUv2 dataset. One can observe that recent works all have very similar performances on this dataset. For example, RDFNet [38] performs 5.3% better than MMAF-Net [16] on NYUv2 dataset, but the gap between them is 0.7% on the SUN RGB-D benchmark. We suspect that this is not only because the SUN RGB-D benchmark is more challenging, but also because it provides more training data that may narrow down the performance gaps between different models. With that being said, our approach still achieves ∼3% absolute performance improvement compared to the previous best number (those without multi-scale testing), which demonstrates that our approach can generalize to a larger dataset.

We also list the main architecture used in different methods in Table 3 for references. According to [32], the complexity (# of parameters and FLOPs) of the Swin-S backbone we use in our framework is similar to a ResNet-101 backbone.

4.5 Qualitative Results

Figure 5 shows the qualitative results of our methods compared to the variant of our method *w/o* UASA and other selected state-of-the-art approaches. We can see that our approach produces decent segmentation results and the quality is consistently better than existing work. We can also observe that adding UASA improves the segmentation results on reflective surfaces such as the glass door

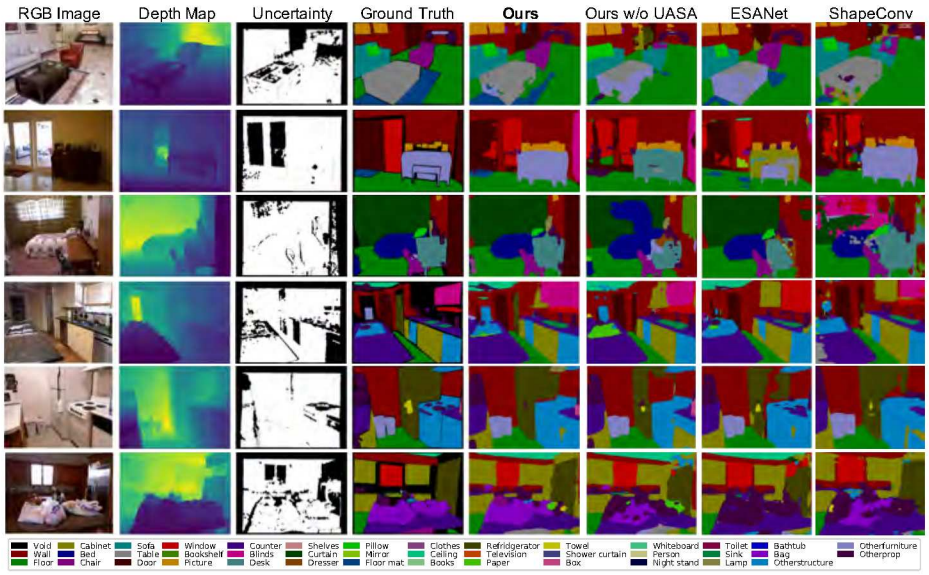

Fig. 5. Qualitative results of our approach compared to baseline and selected existing works.

in row 1 and row 2, the sink in row 4, and the upper part of the refrigerator in row 6. These qualitative results further demonstrate the effectiveness of our proposed method.

5 Conclusion

In this paper, we present a novel framework for indoor RGB-D semantic segmentation. We tackle the two major challenges in this task: 1) how to better extract features from the depth image, and 2) how to effectively fuse and combine information from two modalities. For the first challenge, we propose an Uncertainty-Aware Self-Attention to replace the traditional Self-Attention in a transformer encoder to explicitly control the information flow between uncertain and confident nodes. For the second challenge, we review and analyze the problems of existing fusion modules and design a new fusion module following four design principles. We perform various ablation studies to demonstrate the effectiveness of our proposed methods, and our experimental results show that our approach achieves new state-of-the-art on two public benchmarks with significant improvements compared to existing works.

Acknowledgement: This work was partially supported by a gift from Qualcomm Inc.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7088–7097 (2021)
3. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017)
4. Chen, C., Wei, J., Peng, C., Zhang, W., Qin, H.: Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion. IEEE Transactions on Image Processing **29**, 4296–4307 (2020)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
9. Chen, L.Z., Lin, Z., Wang, Z., Yang, Y.L., Cheng, M.M.: Spatial information guided convolution for real-time rgbd semantic segmentation. IEEE Transactions on Image Processing **30**, 2313–2324 (2021)
10. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 561–577. Springer (2020)
11. Chen, Y., Mensink, T., Gavves, E.: 3d neighborhood convolution: Learning depth-aware features for rgb-d and rgb semantic segmentation. In: 2019 International Conference on 3D Vision (3DV). pp. 173–182. IEEE (2019)
12. Chu, X., Zhang, B., Tian, Z., Wei, X., Xia, H.: Do we really need explicit position encodings for vision transformers? arXiv e-prints pp. arXiv–2102 (2021)
13. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2393–2402 (2018)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)

16. Fooladgar, F., Kasaei, S.: Multi-modal attention-based fusion model for semantic segmentation of rgb-depth images. arXiv preprint arXiv:1912.11691 (2019)
17. Fu, J., Liu, J., Wang, Y., Zhou, J., Wang, C., Lu, H.: Stacked deconvolutional network for semantic segmentation. IEEE Transactions on Image Processing (2019)
18. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3052–3062 (2020)
19. Ghiasi, G., Fowlkes, C.C.: Laplacian pyramid reconstruction and refinement for semantic segmentation. In: European conference on computer vision. pp. 519–534. Springer (2016)
20. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: European conference on computer vision. pp. 345–360. Springer (2014)
21. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv preprint arXiv:2103.00112 (2021)
22. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian conference on computer vision. pp. 213–228. Springer (2016)
23. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3562–3572 (2019)
24. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7519–7528 (2019)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
26. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1440–1444. IEEE (2019)
27. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3d object dataset: Putting the kinect to work. In: Consumer depth cameras for computer vision, pp. 141–165. Springer (2013)
28. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9471–9481 (2021)
29. Jiang, J., Zheng, L., Luo, F., Zhang, Z.: Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv preprint arXiv:1806.01054 (2018)
30. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: ACM SIGGRAPH 2004 Papers, pp. 689–694 (2004)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
33. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

34. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
35. Luo, A., Li, X., Yang, F., Jiao, Z., Cheng, H., Lyu, S.: Cascade graph neural networks for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 346–364. Springer (2020)
36. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1520–1528 (2015)
37. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2022)
38. Park, S.J., Hong, K.S., Lee, S.: Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 4980–4989 (2017)
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32**, 8026–8037 (2019)
40. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7254–7263 (2019)
41. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.M.: Efficient rgb-d semantic segmentation for indoor scene analysis. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13525–13531. IEEE (2021)
42. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
43. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
44. Sun, P., Zhang, W., Wang, H., Li, S., Li, X.: Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1407–1417 (2021)
45. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
46. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. International Journal of Computer Vision **128**(5), 1239–1285 (2020)
47. Wang, W., Neumann, U.: Depth-aware cnn for rgb-d segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 135–150 (2018)
48. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021)
49. Xia, X., Kulis, B.: W-net: A deep model for fully unsupervised image segmentation. arXiv preprint arXiv:1711.08506 (2017)
50. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE international conference on computer vision. pp. 1625–1632 (2013)

51. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 418–434 (2018)
52. Xing, Y., Wang, J., Chen, X., Zeng, G.: 2.5 d convolution for rgb-d semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1410–1414. IEEE (2019)
53. Xing, Y., Wang, J., Chen, X., Zeng, G.: Coupling two-stream rgb-d semantic segmentation network by idempotent mappings. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1850–1854. IEEE (2019)
54. Xing, Y., Wang, J., Zeng, G.: Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 555–571. Springer (2020)
55. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986 (2021)
56. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)