Multi-Modal Attention-based Fusion Model for Semantic Segmentation of RGB-Depth Images

Fahimeh Fooladgar, and Shohreh Kasaei, *Senior Member, IEEE*

Department of Computer Engineering, Sharif University of Technology,Tehran, Iran

The 3D scene understanding is mainly considered as a crucial requirement in computer vision and robotics applications. One of the high-level tasks in 3D scene understanding is semantic segmentation of RGB-Depth images. With the availability of RGB-D cameras, it is desired to improve the accuracy of the scene understanding process by exploiting the depth features along with the appearance features. As depth images are independent of illumination, they can improve the quality of semantic labeling alongside RGB images. Consideration of both common and specific features of these two modalities improves the performance of semantic segmentation. One of the main problems in RGB-Depth semantic segmentation is how to fuse or combine these two modalities to achieve more advantages of each modality while being computationally efficient. Recently, the methods that encounter deep convolutional neural networks have reached the state-of-the-art results by early, late, and middle fusion strategies. In this paper, an efficient encoder-decoder model with the attention-based fusion block is proposed to integrate mutual influences between feature maps of these two modalities. This block explicitly extracts the interdependences among concatenated feature maps of these modalities to exploit more powerful feature maps from RGB-Depth images. The extensive experimental results on three main challenging datasets of NYU-V2, SUN RGB-D, and Stanford 2D-3D-Semantic show that the proposed network outperforms the state-of-the-art models with respect to computational cost as well as model size. Experimental results also illustrate the effectiveness of the proposed lightweight attention-based fusion model in terms of accuracy.

Index Terms—Semantic segmentation, attention-based fusion, multi-modal fusion.

I. INTRODUCTION

SEMANTIC segmentation of RGB-Depth images has been considered as one of the main tasks for 3D scene understanding. The popularity of its applications such as autonomous driving, augmented virtual reality, and the inference of support relations among objects in robotics emphasize the importance of scene understanding. Most of the researches in this field have been done on outdoor scenes which are less challenging compared to indoor scenes. The existence of small objects, light-tailed distribution of objects, occlusions, and poor illumination cause major challenges in indoor scenes, to name a few.

By introducing the Microsoft Kinect camera [1] which captures both RGB and depth images, some indoor semantic segmentation approaches have been concentrated on the RGB-D dataset which alleviates the challenges of the indoor scene. For instance, when RGB images have a poor illumination in some regions, depth images can improve the labeling accuracy. Figure 1 shows some examples in which RGB images have poor lightning in some regions while depth images hold discriminative features.

Utilizing the 3D geometric information in semantic segmentation methods has been provided by the availability of Microsoft Kinect camera [2]. Extracting this 3D geometric information that might be missed in RGB images aids to diminish some uncertainties in dense prediction and object detection processes [3], [4]. Early RGB-D semantic segmentation proposed novel engineering features extracted from RGB and depth images by using intrinsic and extrinsic camera parameters [3], [5], [4], [6], [7]. Then, all of these appearances and 3D features were incorporated into feature vectors fed to common classifiers.

Recently, Deep Convolutional Neural Networks (DCNNs) [8], [9], [10] have improved the accuary of almost all categories of computer vision methods; such as image classification [11], [10], [8], object detection [5], [12], action recognition [13], depth estimation [14], [15], pose estimation [16] , image segmentation [17] and semantic segmentation [18], [19], [20], [21].

Pooling operations and stride convolutions which are applied in CNNs (to become invariant to most local changes) produce the low spatial resolution outputs for dense prediction applications (such as semantic segmentation, depth estimation, and surface normal estimation). Hence, the early deep learning methods [22], [23] for semantic segmentation utilized the deep networks as feature extractors. They then applied a classifier to categorize each pixel, superpixel, or region. Long et al. [24] changed CNNs to Fully Convolutional Neural Networks (FCN) which are more appropriate for dense prediction applications. DeconvNet [25], dilated convolution [26], and unpooling [27] methods have been proposed to recover this spatial information lost. Among these methods, some non-parametric approaches [28], [29] have been proposed where they utilize similarity measurements to label each part of images.

As one of the goals of this paper is the semantic segmentation of RGB-Depth images, the focus is on the main challenges and approaches of RGB-D datasets. The main challenge in RGB-Depth semantic segmentation is how to represent and fuse the RGB and depth channels so that the strong correlations between the depth and photometric channels are considered. Simple methods for fusion of RGB and depth channels are based on the early fusion [22] and late fusion [18] polices.

In this paper, the encoder-decoder architecture with the

Fig. 1: Pairs of RGB and depth images.

novel multi-modal Attention-based Fusion Block (AFB) is proposed to fuse these two modalities in order to obtain more powerful and meaningful RGB-Depth fused feature maps. The attention-based fusion block has been inspired by the attention modules in the squeeze and excitation network [30] which is focused on the channel-wise recalibration of feature maps to model the dependency of channels. The intermediate feature maps extracted from RGB and depth channels of two encoders are considered as the input to the attention-based fusion block. This block computes attention maps which are multiplied by input feature maps for adaptive feature fusion. The attention-based fusion block consists of two sequentially channel- and spatial-wise attention mechanisms to construct the attention maps. Consequently, feature maps of two modalities are fused based on their interdependencies among different channels. Fig 2 illustrates the proposed architecture of attention-based fusion block. Moreover, each AFB is followed by the lightweight chained residual pooling layers to consider the global contextual information in the proposed decoder side. Consequently, the proposed encoder-decoder architecture is an efficient model in terms of the computational cost and the number of parameters.

Main contributions of this work are listed as:

- Proposing an efficient encoder-decoder architecture for semantic segmentation of RGB-Depth images.
- Proposing an attention mechanism of CNNs for modality fusion.
- Incorporating a channel-wise alongside spatial-wise interdependencies for fusion.
- Proposing a novel representation of evaluation metric for semantic segmentation methods.

The remainder of this paper is organized as follows. In Section II, the related work of RGB and RGB-Depth semantic segmentation are categorized. The overall architecture of proposed encoder-decoder model with the fusion block is presented in Section III. The experimental results evaluated on the existing RGB-D dataset by the proposed semantic segmentation criterion are investigated in Section IV. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Traditional approaches of semantic segmentation[3], [2], [4], [6], [7], [5], [31], [32], [33] have two main phases of feature extraction and classification. Engineering or hand-crafted features (such as SIFT, HOG, and SURF) are extracted from pixels, super-pixels, or segmented regions. Then, these features are fed to common classifiers; such as Support Vector Machine (SVM) and Random Forest (RF).

By emerging convolutional neural networks [11], [10], [8], the most successful methods in the field of semantic segmentation have been proposed based on CNNs. Early CNNs methods proposed in semantic segmentation field [22], [34], [23] utilized them as a deep feature extractor. Couprie et al. [22] extracted deep and dense hierarchical features for each region of the segmentation tree [35] via the multi-scale CNN model. These deep features are then fed to the SVM classifier to predict the label of each region.

Well-known CNN models (such as GoogLeNet [36], ResNet [8], [37], and DenseNet [9]) were originally proposed for the image classification task, where the input was an image and the network output was its predicted label. But, these models need some changes to be appropriate for a dense prediction task. The cascaded down-sampling is performed by max or average pooling and then the stride convolution decrease the spatial resolution of feature maps, hence the outputs of these models for the semantic segmentation are very coarse. In other words, the localization information is lost at the end of the networks. The Fully Convolutional Network [18], [24] converted the fully connected layers to the convolutional layers and make them suitable for a dense prediction task; such as semantic segmentation, depth estimation [15], [14], surface normal prediction [15] and video semantic segmentation [38]. They take an image as input and produce corresponding per pixel labeled image in an end-to-end training procedure. Fu et al. [39] proposed Refinet model to improve the FCN method via a segmentation-based pooling idea. The goal of their pooling idea is to maintain the fine-grain details and boundary maps of salient objects.
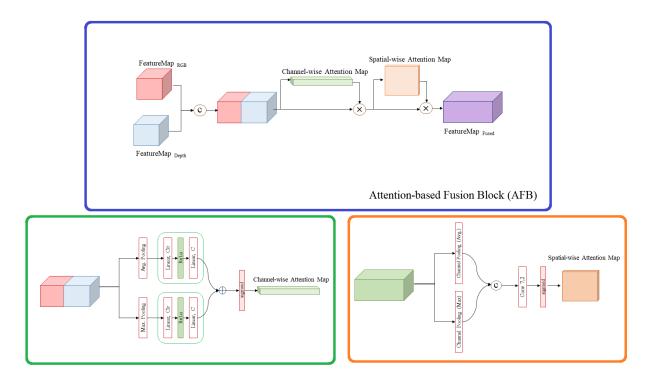
Fig. 2: Attention-based fusion block.

To recover the information loss, different approaches have been proposed. Long et al. [18], [24] up-sampled the feature maps of the last layer and concatenated them with the previous intermediate feature maps in a stage-wise training procedure. The encoder-decoder type models [40], [41], [42] have been proposed to handle the dense per-pixel prediction problem. Commonly, the popular CNN models (such as VGG net, GoogleNet, ResNet, and DenseNet) have been utilized as the encoder to produce intermediate deep feature maps. Then, the goal of decoder branch is to restore the information lost causes in the encoder side.

Different types of approaches have been presented for the decoder side. DeconvNet [25] applied the convolution transposed in the decoder side instead of convolution layers of encoder branch. DeepLab [19], [26], [43] eliminated all max-pooling layers of VGG and applied the dilated convolution to enlarge the receptive field of filters to compensate the effect of max pooling operations. Preserving the index of max-pooling and applying un-pooling operations in the decoder was presented by SegNet model [42]. They proposed an encoder-decoder model in which they utilized the max-pooling indices in the decoder to recover the location of information loss. The fusion of the long-range residual connections of ResNet model was propounded by Lin et al [40] to refine the resolution loss of the CNN architecture.

The main goal of this paper is to address the semantic segmentation challenges of RGB-Depth images. The major issue is how to extract the strong feature representations of both photometric and depth channels. In the case of RGB-Depth semantic segmentation, almost all methods exploit the depth images as another channel of the image. As such, the fusion strategy plays an important role. These strategies can be classified into early, middle, and late fusion. Long et al. [24] proposed the late fusion of FCN while [22] utilized early fusion, and [44] applied middle fusion of RGB and depth channels. The FuseNet [44] considered two encoder branches, one encoder branch for the depth channel and another encoder branch for fusion of RGB and depth channels. Wang et al. [45] designed a transformation block to fuse the common and specific features of the RGB and depth channels of convolution network to bridge the convolutional and deconvolutional models. Li et al. [46] incorporated the vertical and horizontal Long Short-Term Memorized (LSTM) method to exploit the interior 2D global contextual relations of RGB and depth channels, separately. Then, the horizontal LSTM has been applied to their concatenated feature maps. Liu et al. in [47] improved the HHA coding of [7] via integrating 2D and 3D information. They then extended the VGG [10] architecture proposed by [19] for RGB-D semantic segmentation. They proposed the weighted summation of RGB and depth streams of a CNN model followed by a fully connected CRF to enhance the prediction.

The RDFNet proposed by Park et al. [48] extended the RefineNet [40] for RGB-Depth images. They considered two encoder streams (RGB encoder and depth encoder), one fusion stream and one decoder stream. They utilized the cascaded refinement blocks of the RefineNet as their decoder stream. The refinement process was applied to the fusion of the RGB and depth feature maps to emend the resolution loss. The Multi-Modal Multi-Resolution RefineNet (3M2RNet) [49] proposed the fusion of long-range residual connections of two ResNet encoder branches with focus on the identity mapping idea
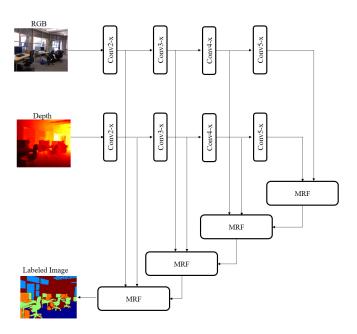
Fig. 3: Proposed network architecture.

Fig. 4: Building block of a residual unit[8].

of RefineNet [40]. Lin et al. [50] proposed a context-aware receptive field based on the scene resolution to incorporate the relevant contextual information. Consequently, for each scene resolution, deep features were learned specifically in a cascaded manner to exploit the relevant information of the neighborhood. Depth-aware convolution and pooling operations were presented by [51] to investigate the 3D geometry of the depth channel. Kang et al. [52] have proposed the depth-adaptive receptive field in the fully convolutional neural network where the size of each receptive filed has been selected based on the distance of each point from the camera. Hence, they utilize the depth images to determine the size of each filter for each neuron adaptively.

III. PROPOSED METHOD

The proposed Multi-Modal Attention Fusion Network (MMAF-Net) model is an encoder-decoder CNN architecture with two simultaneously encoder branches of RGB and depth modalities as inputs while including one decoder branch. Both of the encoder branches follow the structure of residual block proposed in the ResNet model [8]. In the decoder branch, the feature maps of both encoder branches from the same level of resolution have been fused based on the novel proposed attention fusion modules to combine both appearance and 3D feature maps. These fused feature maps have been utilized to recover the information loss of encoders and produce a high resolution prediction output. The overall view of proposed encoder-decoder architecture is illustrated in Figure 3. In the following subsections, the proposed encoder-decoder architecture with a more focus on the multi-modal multi-resolution fusion block of the proposed decoder stream is explained.

A. Encoder Stream

All of the well-known CNN models were primarily proposed for image classification. In those networks, at the end of
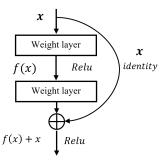
networks the high semantic but low spatial resolution feature maps produce rough segmentation results for semantic segmentation purposes. To overcome this limitation, an encoder-decoder model is proposed. In the encoder part of the proposed model, the residual blocks of the ResNet model are utilized to benefit from the short and long range skip connections properties. The short-range skip connections immune the networks from the vanishing gradients problem while the long-range skip connections help to refine the information loss caused by the cascaded down-sampling operations and stride convolutions.

As illustrated in Figure 3, the proposed model utilizes the residual blocks of ResNet model (Convi-x) as two separate encoder branches. He et al. in [37], analyzed and compared the rule of skip connections of the scaling, gating, 1×1 convolution, and the identity mapping. They showed that using the identity mapping function (as the skip connection) in a deep residual network is more helpful for the generalization of the network as well as the convergence of the optimization algorithm. The building block of one residual unit is illustrated in Figure 4. This short-range skip connection can be formulated as $y_l = F(x_l, W_l) + H(x_l)$, where x_l is the original input to the residual unit l, and $x_{l+1} = G(y_l)$ is the output of unit l, which is fed to the residual unit of $l + 1$ as its input. Also, $F(x_l, W_l)$, $H(x_l)$, and $G(y_l)$ are the series of the operations applied on the input x_l or y_l (such as convolution, batch normalization, and nonlinearity). In the first version of the ResNet model [8], H is set as an identity function and the Relu function is used for G. Therefore, the information flow of x_l is not changed and added by $F(x_l, W_l)$. Hence, this residual unit with the identity and Relu functions enhances the performance of very deep networks as the vanishing gradient problem was solved. Consequently, if in the encoder-decoder model the weights of network are very small, the gradient of layers does not completely vanish. Thereupon, the vanishing gradients problem does not occur in such a deep network.

Between each residual block of the ResNet, the sequential down-sampling operations (applied by the pooling layers) increase the receptive field of the filters to include more context and also prevent the growth in the number of training weights through the encoder stream. Therefore, they preserve efficient and tractable training. But, the network loses some valuable information. This information loss produces the low-resolution prediction in the dense per-pixel classification in

which the localization of the semantic labels is more essential than the image classification applications. This means that the higher-level feature maps of deeper layers in the multi-encoders which encode the high-level semantic information and carry more object-level information suffer from the lack of localization information. Here, it is proposed to recover this information loss in the up-sampling process of the decoder branch by the attention-based fusion of the long-range residual connections of multi-encoder streams with the preceding decoder output. Therefore, the decoder part is responsible to recover this resolution loss in cascaded multi-modal multi-resolution fusion blocks.

B. Decoder Stream

The proposed model applies efficient multi-modal attention-based fusion modules in the decoder branch of the network to recover the information loss caused by the down-sampling processes in multi encoder streams. The goal of the decoder is to employ the multi-level feature maps coming from the long-range skip connections of two encoder branches to enhance the resolution which is lost by the down-sampling operation performed by the pooling or convolution layers (with $stride > 1$).

The output of residual blocks of encoder branches are employed as the long-range skip connections and are fed to the 4-cascaded sub-modules of the decoder, called Multi-Modal Multi-Resolution Fusion (MRF) module. As such, it actually utilizes long and short residual connections. These skip connections, along with the attention-based fusion modules, enable efficient end-to-end training of RGB-Depth encoder-decoder model as well as the efficient high-resolution prediction.

The overall structure of the MRF module with three modalities as its inputs is illustrated in Figure 5. The proposed decoder has 4 cascaded MRF modules. It consists of two main sub-blocks of Attention Fusion Module (AFM), and a Chained Residual Pooling (CRP). It has three input modalities including: i) feature maps extracted from the RGB encoder branch, ii) feature maps extracted from the depth encoder branch at the same resolution level, and iii) the feature maps of the preceding MRF at the lower resolution. In the AFM, two fusion policies have been performed. The first one is the AFB, where the attention-based fusion strategy is applied to two first inputs of this module. The second one is a simple summation strategy to fuse the output of the previous MRF with the output of the current AFB to perform the refinement and produce the high resolution feature maps. The idea of CRP sub-block is to capture the context in multiple region sizes with the chained residual pooling layers.

1) Attention-based Fusion Block

The attention mechanism in deep convolutional neural networks is based on visual attention which is consistent with the human visual perception. To perceive the scene and object structure, human visual system focuses on the salient parts. In fact, it concentrates on the most noticeable or important parts of the scene in different sequential glances. We propose to investigate this attention for modality fusion to focus on the salient parts of feature maps in each modality. Recently, different attention-based and salient object detection [53], [54], [55] have been proposed where their goal is to model the human attention mechanism. To the best of our knowledge, no previous work has investigated the attention strategy for modality fusion.

The proposed AFB block sequentially considers channel-wise and spatial-wise attention. The goal of channel-wise attention is to determine salient channels of concatenated feature maps, while spatial-wise attention denotes "where" salient feature maps are located.

The proposed encoder architecture consists of four sub-blocks with the convolution, non-linearity function, and down-sampling operation. Hence, each encoder branch produces four intermediate feature maps. Intermediate feature maps of the same level in two encoder branches are concatenated and fed to the corresponding AFB in their level.

The structure of the proposed AFB is inspired by the Convolutional Block Attention Module (CBAM) of [56]. They proposed two sequential channel and spatial attention modules to refine the intermediate extracted feature maps. They illustrated that this module integrates the focus of the network to the target object in an image.

The goal of our attention-based fusion is to enhance the representation power of concatenated RGB-Depth feature maps and capture their salient feature maps while suppressing the unnecessary ones. The intermediate feature maps of each encoder stream demonstrate a set of local descriptors where their statistics can be considered as a good representative for each image. These statistics include the average and maximum of each feature map. In the proposed fusion method, the non-linear and non-mutually exclusive relationships between RGB and Depth intermediate feature maps are exploited via the pooling, non-linearity, convolution, and fully connected layers of the deep neural networks operations.

Suppose $F_{RGB} \in \mathbb{R}^{n \times m \times c}$ and $F_D \in \mathbb{R}^{n \times m \times c}$ are intermediate feature maps of RGB and Depth modalitis in the same level, respectively, and $F = [F_{RGB}; F_D] \in \mathbb{R}^{n \times m \times 2c}$ shows their concatenation. The channel-wise attention map, M_c, is computed as

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$
$$= \sigma(W_1(W_0(AvgPool(F))) + W_1(W_0(MaxPool(F)))) \quad (1)$$

where σ denotes the sigmoid function. Then, $F^{'} \in \mathbb{R}^{n \times m \times 2c}$ is determined as an output of the channel-wise attention module ($F^{'} = M_c(F) \times F$). The spatial-wise attention map, M_s, is applied on $F^{'}$ and is computed as

$$M_s(F^{'}) = \sigma(Conv([AvgPool(F^{'}); MaxPool(F^{'})])). \quad (2)$$

Then, $F^{''} \in \mathbb{R}^{n \times m \times 2c}$ is determined as an output of the spatial-wise attention module ($F^{''} = M_s(F^{'}) \times F^{'}$). Consequently, $F^{fused} \in \mathbb{R}^{n \times m \times c}$ is determined by an output of the spatial-wise attention module ($F^{fused} = MaxPool3D(F^{''})$). This fused feature map focuses on the important features of the channels and concentrates on "where" salient features are located.
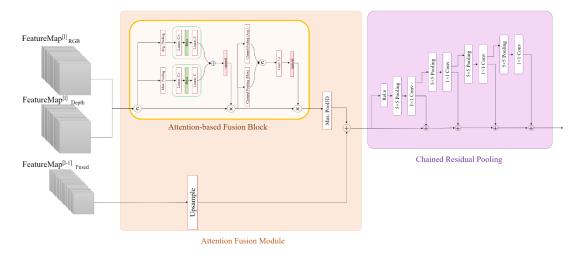
Fig. 5: Proposed multi-modal multi-resolution fusion module.

2) Chained Residual Pooling

To efficiently recover the information loss of encoder feature maps, the contextual information is exploited from large regions of an image by utilizing the cascaded pooling operations in a chain residual connections (see Figure 5). The 4-cascaded $5 \times 5 Pooling - 1 \times 1 Conv$ layer can capture long-range contextual information with a fixed pooling window size. These pooling feature maps combine with each other via the learnable 1×1 filters of convolutional layers. Consequently, all of these pooling feature maps are located in residual connections; hence each output fuses with input feature maps by a simple summation operation.

C. Training Procedures of MMAF

To learn the proposed network, the training set is defined as

$$\{(X_i^{RGB}, X_i^D, Y_i)|X_i^{RGB} \in \mathbb{R}^{(H \times W \times 3)}, X_i^D \in \mathbb{R}^{(H \times W)},$$
$$Y_i \in L^{(H \times W)}, i = 1, , N\}$$

where L denotes the labeling set defined as $L = 1, , C$ in which C and N determine the number of class labels and training data, respectively. The output of the networks is considered as the function $f(x^{RGB}, x^D; \mathbb{W})$ that is the composition of functions corresponding to each network layer and $\mathbb{W}$ denotes all of the network weights. The probability of a pixel x for a given class c with the soft-max function is computed as

$$p(\hat{y} = c|x^{RGB}, x^D, \mathbb{W}) = \frac{e^{(f_c(x^{RGB}, x^D; \mathbb{W}))}}{\sum_{l=1}^{C} e^{(f_l(x^{RGB}, x^D; \mathbb{W}))}}. \quad (3)$$

The simple categorical cross-entropy for the loss function is defined as

$$J = \frac{-1}{M} \sum_{i=1}^{M} \boldsymbol{\ell}(y_i, \hat{y}_i) \quad (4)$$

where $\boldsymbol{\ell}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log p(\hat{y} = c|x_i^{RGB}, x_i^D, \mathbb{W})$. M is the total number of pixels in the training data and y is

the one-hot encoding vector of length C that determines the ground-truth label of pixel i.

IV. EXPERIMENTAL RESULTS

This section contains two main subsections. In the first one, the proposed attention-based fusion method is evaluated on the challenging SUN-RGBD [57], NYU-V2 [4], and Stanford-2D-3D-Semantic [58] datasets. These three datasets contain RGB and depth images with the corresponding dense per pixel ground-truth (GT) images. Then, in the second subsection, the evaluation metrics of semantic segmentation are perused to find a more proper approach to analyze prediction outputs of each model on each dataset.

A. Attention-based Fusion Results

To compare the efficiency of the proposed method, it has been compared with the state-of-the-art CNN models. It is implemented using the PyTorch library. The random cropping, scaling, and flipping operations have been utilized for data augmentation. The weights of ResNet model are employed as the pre-trained weights of two encoder branches. The categorical cross-entropy has been considered as the loss function which is optimized by the Stochastic Gradient Descent (SGD) algorithm. The global accuracy (G), the mean accuracy (M), and the Intersection-over-Union (IoU) score have been computed for evaluation purposes.

1) Evaluation Results on SUN-RGBD Dataset

SUNRGB-D dataset contains 10335 RGB-Depth images as well as dense per-pixel labeling with specific train and test set splits. The dataset has assigned each pixel into one of 37 valid classes, where one class label has been assigned to void. It is worth mentioning that the distribution of labels in this dataset is considerably unbalanced and it is remarkable that approximately 25% of pixels in the training data have not been assigned to any of 37 valid classes and are set as the void class.

The experiments were performed on three different ResNet models as two encoder streams. To investigate the performance of the proposed multi-modal multi-resolution fusion

modules, the performance of the method is compared with the simple middle fusion strategy (SMF-Net) as well as the single modality of RGB and Depth channel. The SMF-Net is the proposed model without attention-based fusion block. In fact, its attention-based fusion block has been replaced with a simple summation operation. As reported in Table I, the attention fusion policy has attained higher accuracies by adding less than $0.5M$ parameters. For comparison purposes, the results of the model with the single encoder branch of RGB or depth modality are also reported in Table I .

The performance of the proposed method is compared with some previous approaches that utilize RGB and depth channels. As summarized in Table II, the proposed method has achieved a higher mean and global accuracy as well as IoU score than other well-known methods that utilize just RGB images (such as SegNet, DeepLab, RefineNet, and some others). In comparison with the state-of-the-art methods applied on RGB-D images, the proposed fusion method attains better results than the FuseNet, LSTM-CF, and D-CNN while it achieves comparable accuracy with the RDFNet and 3M2RNet methods.

The proposed multi-modal multi-resolution fusion method is more computationally efficient than these two methods. In Table III, the proposed method has been compared with them based on computational complexity and model size. The proposed method is more advisable for applications running on embedded devices or even those that require real-time performance. The RDFNet model did not report its computational cost and model size. But, it is an extension of the RefineNet for RGB-Depth images. Hence, it has inevitably more parameters and computational cost than the RefineNet, because it has one extra encoder stream for depth channel and one additional fusion stream to fuse RGB and depth feature maps. It is also notable that the accuracy of RefineNet, RDFNet, and 3M2RNet model have been reported based on a multi-scale evaluation where all of the other accuracies are reported on a single-scale evaluation.

In Figure 8, some test samples of the SUN-RGBD dataset that are predicted by the proposed method are depicted.

2) Evaluation Results on NYU-V2 Dataset

The NYU-V2 dataset is known as the most popular dataset among indoor RGB-D datasets. It contains images from 646 different scenes with 26 variants of scene types. It includes 1449 RGB and depth images with per-pixel annotation which are splitted to 795 training images and 654 test images. Their class labels are mapped to 40 class labels by Gupta et al. in [6]. The dataset is unbalanced with respect to the ratio of the number of pixels per class objects and contains the label void showing the pixels which cannot be annotated.

The proposed method has been compared with the most important CNN models. As the results listed in Table IV show, it has surpassed all of CNN models using a single RGB modality. For instance, it obtained approximately %6 higher mean IoU than the Context model [59]. It has also achieved better results than methods that have utilized both RGB and depth channels, while the MMAF-Net did not outperform the RDFNet and 3M2RNet model in terms of accuracy. But, note that it has a lower model size as well as computational

complexity than these two models (see Table III). These two models, as well as the RefineNet, have evaluated their results based on the multi-scale method (determines with '*' in Table IV).

3) Evaluation Results on Stanford-2D-3D-Semantic Dataset

It contains 70496 RGB and depth images as well as 2D annotation with 13 object categories. It includes 1413 RGB and depth panoramic images as well as their surface normal and semantic annotations of six large-scale indoor areas. It also provides 3D point clouds of these areas. Areas 1, 2, 3, 4, and 6 are utilized as the training and Area 5 is used as the testing set. The attention-based fusion model was applied on RGB-D images (not panoramic ones). Hence, Table V shows the performance comparison with those approaches that have been evaluated on RGB-D images. The authors of the D-CNN model [51], evaluated their model as well as the DeepLab [19] model on this dataset. They trained these two models from scratch. The proposed MMAF-Net has obtained comparable performance with the 3M2RNet model in terms of accuracy while it enjoys a lower model size and computational complexity (see Table III). Tateno et al. [62] and Kong et al. [63] reported their accuracy on the panoramic images of this dataset.

B. Proposed Evaluation Metrics for Semantic Segmentation

The semantic segmentation problem is actually known as a dense labeling problem. Hence, evaluation metrics that had been utilized for labeling in the machine learning field have also been applied to semantic segmentation methods. Accordingly, the confusion matrix has been computed and then the global, mean, and IoU criteria have been figured out from it. There are two main issues related to these criteria used for semantic segmentation methods which are explained in the following with more details.

First issue: Almost all of the semantic segmentation approaches calculate these metrics per pixel for whole images of each dataset (per dataset). For example, the $global\ accuracy = 81$ means %81 of all pixels of all test images have been classified correctly. It does not carry out any additional information about each image's accuracy. Hence, all pixels of one test image may be classified correctly but the other ones may have a large misclassification error. As a result, it is ambiguous whether the method has approximately the same performance for all images or it has a rich performance for some of them and a poor performance for others. Csurka et al. [64] proposed to measure the per image accuracy instead of per dataset. They computed the confusion matrix for each image based on the union of classes presented in the ground-truth as well as in the prediction. Therefore, the number of images that have attained an accuracy more than a specific threshold can be reported. But, almost all of the existing methods followed the former accuracy metrics which are computed per dataset. We propose to compute the global, mean, and IoU metrics for each test image, separately and depict their Cumulative Distribution Function (CDF) to illustrate the least number of images with a specified accuracy

TABLE I: Performance evaluation of proposed MMAF module on SUN-RGBD dataset.

Methods	Modality	G	M	IoU	W-IoU	No. of Parameters	GFLOPs
MMAF-Net-50 (ours)	RGB	78.4	53.1	42.3	65.9	27.4M	32.7G
MMAF-Net-50 (ours)	D	74.8	44.7	35.4	61.3	27.4M	32.7G
SMF-Net-50 (ours)	RGB-D	79.0	55.2	43.7	67.0	52.6M	56.7G - 464.5K
MMAF-Net-50 (ours)	RGB-D	80.0	57.6	45.5	68.0	53.0M	56.7G
MMAF-Net-101 (ours)	RGB-D	80.2	58.0	46.0	69.0	91.0M	95.6G
MMAF-Net-152 (ours)	RGB-D	81.0	58.2	47.0	69.6	122.3M	134.4G

TABLE II: Semantic segmentation results on SUN RGB-D dataset ('*' denotes multi-scale evalution).

Methods	Modality	G	M	IoU
Ren et al. [3]	RGB-D	-	36.3	-
DeconvNet [25]	RGB	66.1	32.3	22.6
FCN [18]	RGB	68.2	38.4	27.4
SegNet [27]	RGB	72.6	44.8	31.8
B-SegNet [41]	RGB	71.2	45.9	30.7
DeepLab [26]	RGB	71.9	42.2	32.1
LSTM-CF [46]	RGB-D	-	48.1	-
FuseNet [44]	RGB-D	76.3	48.3	37.3
Context [59]	RGB	78.4	53.4	42.3
D-CNN [51]	RGB-D	-	53.5	42.0
3D Graph [60]	RGB-D	-	57.0	45.9
Cheng et al. [61]	RGB-D	-	58.0	-
RefineNet [40]	RGB	80.6*	58.5*	45.9*
CFN (VGG-16)[50]	RGB-D	-	-	42.5*
CFN (RefineNet)[50]	RGB-D	-	-	48.1*
RDFNet [48]	RGB-D	81.5*	60.1*	47.7*
3M2RNet [49]	RGB-D	83.1*	63.5*	49.8*
MMAF-Net-152 (ours)	RGB-D	81.0	58.2	47.0

TABLE III: Computational complexity and model size comparison on SUN RGB-D dataset ('*' denotes multi-scale evalution).

Methods	G	M	IoU	Parameters	GFLOPs
RefineNet [40]	80.6*	58.5*	45.9*	119.0M	234.9G
3M2RNET [49]	83.1*	63.5*	49.8*	225.4M	384.5G
RDFNET [48]	81.5*	60.1*	47.7*	-	-
MMAF-Net (ours)	81.0	58.2	47.0	122.3M	134.4G

TABLE IV: Semantic segmentation results on NYU-V2 dataset ('*' denotes multi-scale evalution).

Methods	Modality	G.	M	IoU
Silberman et al. [4]	RGB-D	54.6	19.0	-
Ren et al. [3]	RGB-D	49.3	21.1	21.4
Gupta et al. [6]	RGB-D	59.1	28.4	29.1
Gupta et al. [7]	RGB-D	60.3	35.1	31.3
Eigen et al. [15]	RGB	65.6	45.1	34.1
FCN [18]	RGB-D	65.4	46.1	34.0
Wang et al. [45]	RGB-D	-	47.3	-
Liu et al. [47]	RGB-D	70.3	51.7	41.2
Context [59]	RGB	70.0	53.6	40.6
Kang et al. [52]	RGB-D	68.4	49.0	37.6
LSTM-CF [46]	RGB-D	-	49.4	-
3D Graph [60]	RGB-D	-	55.7	43.1
D-CNN [51]	RGB-D	-	56.3	43.9
Cheng et al. [61]	RGB-D	71.9	60.0	45.9
RefineNet [40]	RGB	73.6*	58.9*	46.5*
CFN (VGG-16) [50]	RGB-D	-	-	41.7*
CFN (RefineNet)[50]	RGB-D	-	-	47.7*
RDFNet [48]	RGB-D	76.0*	62.8*	50.1*
3M2RNet [49]	RGB-D	76.0*	63.0 *	48.0*
MMAF-Net-152 (ours)	RGB-D	72.2	59.2	44.8

TABLE V: Semantic segmentation results on Stanford-2D-3D-Semantic dataset ('*' denotes multi-scale evalution).

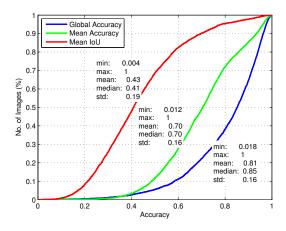| Methods	Modality	G.	M	IoU
DeepLab [19]	RGB-D	64.3	46.7	35.5
D-CNN [51]	RGB-D	65.4	55.5	39.5
3M2RNet [49]	RGB-D	79.8*	75.2 *	63.0*
MMAF-Net-152 (ours)	RGB-D	76.5	62.3	52.9

Fig. 6: Semantic segmentation criteria computed per image presented via CDF.

level. Figure 6 shows this CDF for these three famous criteria. The horizontal axis shows the accuracy. Hence, for instance, approximately 70 percent of test images have more than $\%80$ global accuracy. For each CDF, minimum, maximum, median, mean, and standard deviation have been reported.

Second issue: Region boundaries are one of the main criteria to determine the quality of image segmentation. The accuracy of these boundaries has not been considered by the global, mean, and IoU metrics, separately. Here, the Boundary Displacement Error (BDE) [65] was utilized to measure the average displacement error between two segmented boundaries of two images. For each boundary pixel, this error is defined as the distance between the closest pixel in the other boundary image. This metric has been presented for image segmentation where here we propose to utilize it for each segmented region that belongs to the same class label in the ground-truth and the prediction image. Suppose B^P is the boundary points of a region with class label l in a prediction image and B^G is its corresponding boundary points in the ground-truth image. The two distance distributions are computed from B^G to B^P and from B^P to B^G. Then, the minimum distance of each point of B^P from B^G is considered as $d(x, B^G) = min\{d_E(x,y)\} \, \forall \, y \, in \, B^P$, where d_E is an Euclidean distance.

To apply this metric in semantic segmentation, the BDE is computed for each class label, separately. Figure 7 illustrates the CDF of BDE for each class label. For instance, %60 of images have less than 10 pixels discrepancy for Floor and Chair classes (see Figure 7).

V. CONCLUSION

An efficient attention-based fusion method for RGB and depth fusion was proposed. The proposed method focused on salient feature maps generated from RGB and depth encoder branches and suppressed unnecessary ones to efficiently fuse these two modalities. The network model was a type of encoder-decoder CNN architectures with two encoder branches and one decoder. The decoder goal was to refine the resolution loss caused by the down sampling procedures in encoder branches via fusion of long-range residual connections coming from both encoder branches. The proposed architecture achieved approximately comparable accuracy in terms of the IoU score, mean accuracy, and global accuracy, with RGB-D state-of-the-art methods. This same level of accuracy attained remarkably with %50 better performance in terms of model size alongside less computational cost (approximately 250G less floating points operations).

REFERENCES

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[2] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 601–608.

[3] X. Ren, L. Bo, and D. Fox, "Rgb-(d) scene labeling: Features and algorithms," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2759–2766.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[5] C. Cadena and J. Košecka, "Semantic parsing for priming object detection in rgb-d scenes," in *3rd Workshop on Semantic Perception, Mapping and Exploration*. Citeseer, 2013.

[6] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 564–571.

[7] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[13] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[14] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[15] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[16] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[17] Q. Wang, C. Yuan, and Y. Liu, "Learning deep conditional neural network for image segmentation," *IEEE Transactions on Multimedia*, 2019.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[19] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *International Conference on Learning Representations*, 2015.

[20] H. Shi, H. Li, F. Meng, Q. Wu, L. Xu, and K. N. Ngan, "Hierarchical parsing net: Semantic scene parsing from global scene to objects," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2670–2682, 2018.

[21] Y. Li, Y. Guo, J. Guo, Z. Ma, X. Kong, and Q. Liu, "Joint crf and locality-consistent dictionary learning for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 875–886, 2018.

[22] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[23] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[24] E. S. . J. L. . T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[25] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[28] L. Khelifi and M. Mignotte, "Mc-ssm: Nonparametric semantic image segmentation with the icm algorithm," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 21, no. 8, pp. 1946–1959, 2019.

[29] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Scene parsing with integration of parametric and non-parametric models," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2379–2391, 2016.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[31] A. C. Müller and S. Behnke, "Learning depth-sensitive conditional random fields for semantic segmentation of rgb-d images," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. Citeseer, 2014, pp. 6232–6237.

[32] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *European conference on computer vision*. Springer, 2006, pp. 1–15.

[33] F. Fooladgar and S. Kasaei, "Semantic segmentation of rgb-d images using 3d and local neighbouring features," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, 2015, pp. 1–7.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[35] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
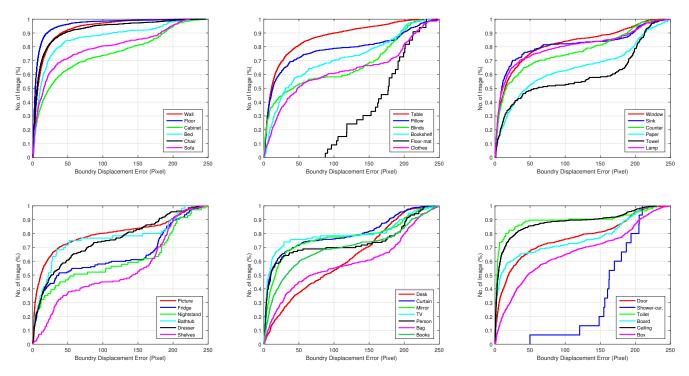
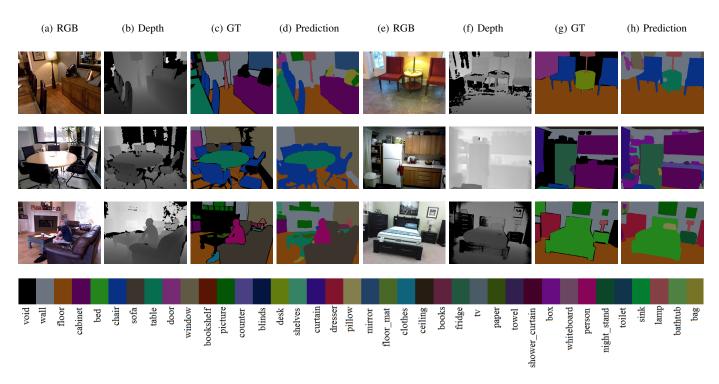Fig. 7: Comparison of boundary displacement error for each class label.

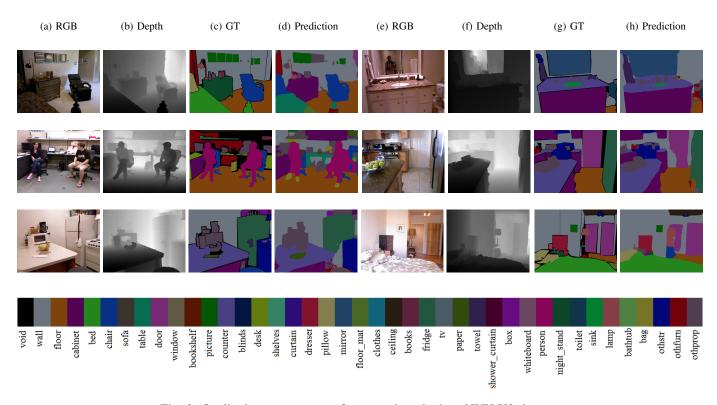Fig. 8: Qualitative assessments of proposed method on SUN RGB-D dataset.

(a) RGB (b) Depth (c) GT (d) Prediction (e) RGB (f) Depth (g) GT (h) Prediction

void wall floor cabinet bed chair sofa table door window bookshelf picture counter blinds desk shelves curtain dresser pillow mirror floor_mat clothes ceiling books fridge tv paper towel shower_curtain box whiteboard person night_stand toilet sink lamp bathtub bag othstr othfurn othprop

Fig. 9: Qualitative assessments of proposed method on NYU-V2 dataset.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[38] Z. Qiu, T. Yao, and T. Mei, "Learning deep spatio-temporal dependence for semantic video segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 939–949, 2017.

[39] K. Fu, Q. Zhao, and I. Y.-H. Gu, "Refinet: A deep segmentation assisted refinement network for salient object detection," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 457–469, 2018.

[40] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." in *Cvpr*, vol. 1, no. 2, 2017, p. 5.

[41] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[42] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[44] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 213–228.

[45] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 664–679.

[46] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling," in *European Conference on Computer Vision*. Springer, 2016, pp. 541–557.

[47] H. Liu, W. Wu, X. Wang, and Y. Qian, "Rgb-d joint modelling with scene geometric information for indoor semantic segmentation," *Multimedia Tools and Applications*, pp. 1–14, 2018.

[48] S.-J. Park, K.-S. Hong, and S. Lee, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[49] F. Fooladgar and S. Kasaei, "3m2rnet: Multi-modal multi-resolution refinement network for semantic segmentation," in *Science and Information Conference*. Springer, 2019, pp. 544–557.

[50] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, "Cascaded feature network for semantic segmentation of rgb-d images," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1320–1328.

[51] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

[52] B. Kang, Y. Lee, and T. Q. Nguyen, "Depth-adaptive deep neural network for semantic segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2478–2490, 2018.

[53] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "Gla: Global–local attention for image description," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 726–737, 2017.

[54] R. Quan, J. Han, D. Zhang, F. Nie, X. Qian, and X. Li, "Unsupervised salient object detection via inferring from imperfect saliency models," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1101–1112, 2017.

[55] K. Fu, I. Y.-H. Gu, and J. Yang, "Saliency detection by fully learning a continuous conditional random field," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1531–1544, 2017.

[56] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[57] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 567–576.

[58] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[59] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1352–1366, 2018.

[60] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5199–5208.

[61] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Localitysensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2017.

[62] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[63] S. Kong and C. Fowlkes, "Pixel-wise attentional gating for parsimonious pixel labeling," *arXiv preprint arXiv:1805.01556*, 2018.

[64] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?." in *BMVC*, vol. 27. Citeseer, 2013, p. 2013.

[65] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *European Conference on Computer Vision*. Springer, 2002, pp. 408–422.