This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

RDFNet: RGB-D Multi-level Residual Feature Fusion for Indoor Semantic Segmentation

Seong-Jin Park
POSTECH

Ki-Sang Hong
POSTECH

Seungyong Lee
POSTECH

{windray,hongks,leesy}@postech.ac.kr

Abstract

In multi-class indoor semantic segmentation using RGB-D data, it has been shown that incorporating depth feature into RGB feature is helpful to improve segmentation accuracy. However, previous studies have not fully exploited the potentials of multi-modal feature fusion, e.g., simply concatenating RGB and depth features or averaging RGB and depth score maps. To learn the optimal fusion of multi-modal features, this paper presents a novel network that extends the core idea of residual learning to RGB-D semantic segmentation. Our network effectively captures multi-level RGB-D CNN features by including multi-modal feature fusion blocks and multi-level feature refinement blocks. Feature fusion blocks learn residual RGB and depth features and their combinations to fully exploit the complementary characteristics of RGB and depth data. Feature refinement blocks learn the combination of fused features from multiple levels to enable high-resolution prediction. Our network can efficiently train discriminative multi-level features from each modality end-to-end by taking full advantage of skip-connections. Our comprehensive experiments demonstrate that the proposed architecture achieves the state-of-the-art accuracy on two challenging RGB-D indoor datasets, NYUDv2 and SUN RGB-D.

1. Introduction

Semantic segmentation that assigns all pixels into different semantic classes is a fundamental task for visual scene understanding. In the past, there was broad research for semantic segmentation based on conditional random field (CRF) using conventional hand-crafted visual features [34, 23, 41]. Recently, deep convolutional neural networks (DCNNs) have achieved great success in image classification task [22, 43, 36, 14]. Built on the success of image recognition using DCNNs, many semantic segmentation methods have also adopted DCNNs by extending them to fully convolutional pixel-wise classification [30, 4, 42].

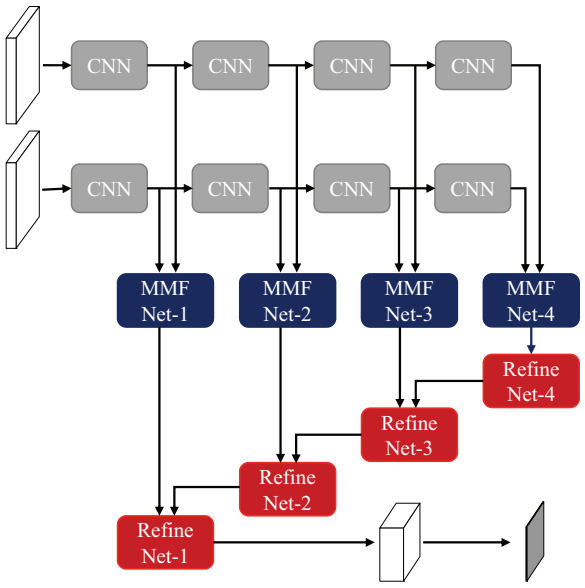

Figure 1. Diagram of the proposed RDFNet for RGB-D semantic segmentation. The network fistly fuses multi-modal features through a block called MMFNet and refines the fused features through a series of RefineNet blocks.

Subsequent research [45, 29, 1, 3, 28] that incorporates the CRF framework into DCNN further improved the accuracy. However, indoor semantic segmentation is still one of the most challenging problem due to complex and various object configurations with severe occlusions.

With the availability of commercial RGB-D sensors such as Microsoft Kinect [44], it has been consistently proved that utilizing features extracted from depth information is useful to reduce the uncertainty for recognizing objects [32, 10, 20, 35, 6, 5, 11, 7, 25, 39, 13]. Depth features can describe 3D geometric information which might be missed in RGB-only features. To extract useful features from both RGB and depth data, it is crucial to develop an effective method for fusing two modalities. There have been many

attempts to utilize the depth information for semantic segmentation in different ways.

Previously most methods [32, 10, 20, 35, 6] designed hand-crafted depth features and constructed various models to classify each region or pixel. In contrast, recent approaches [5, 11, 7, 25, 39, 13] employ DCNNs which successfully learn informative RGB features from low level primitives for high level semantics. As the main issue of RGB-D semantic segmentation is how to effectively extract and fuse depth features along with color features, various approaches have been proposed to exploit the ability of DCNN for integrating depth information. The approaches include concatenating input RGB and D channels, fusing score maps computed from each modality, extracting common and specific features for different modalities, and so on. Although previous approaches achieved meaningful results, there has been a lack of research that fully utilizes recent successful CNN architectures using skip-connections.

In the case of RGB semantic segmentation, Lin *et al*. [26] has recently achieved great success in utilizing multi-level RGB features with different resolutions by iteratively fusing and refining them. They designed a network called *RefineNet* by taking advantage of residual learning with skip-connection [14, 15] that enables effortless backpropagation of gradients during training. The multi-level features in RefineNet are connected through the short and long-range residual connections and thus can be efficiently trained and merged into a high-resolution feature map.

Inspired by the work, we present a novel RGB-D fusion network (RDFNet) that extends the core idea of residual learning to RGB-D semantic segmentation. We extend the RefineNet to effectively extract and fuse RGB and depth features through residual feature fusion. Our network consists of two feature fusion blocks: multi-modal feature fusion (MMF) block and multi-level feature refinement (Refine) block (Figure 1). The MMF block is crucial to exploit different modality of RGB and depth features. The block is constructed by mimicking the RefineNet block but with different inputs; The inputs are multi-level RGB and depth features computed from deep residual network [14]. Then, it fuses the different modality features through residual convolutional units and feature adaptation convolution, followed by optional residual pooling. The MMF block adaptively trains residual feature to effectively fuse the complementary features in different modalities, while learning the relative importance of each modality feature. The block is subsequently followed by the Refine block to further process the fused features for high-resolution semantic segmentation. In this architecture, discriminative multi-level RGB and depth features can be effectively trained and fused, while retaining the key advantage of the skip connection, i.e., all the gradients effectively flow backwards through residual connections to the ResNet input features.

Our main contributions can be summarized as follows:

1. We propose a network that effectively extracts and fuses multi-level RGB-D features in very deep network by extending the core idea of residual learning to RGB-D semantic segmentation.

2. Our multi-modal feature fusion block enables efficient end-to-end training of discriminative RGB-D features on a single GPU by taking full advantage of residual learning with skip-connection.

3. We show that our network for RGB-D semantic segmentation outperforms existing methods and achieves the state-of-the-art performance on two public RGB-D datasets, NYUDv2 and SUN RGB-D.

2. Related Work

Since great advance in image classification task using DCNN [22, 43, 36, 14], most recent semantic segmentation methods have employed DCNN. Long *et al*. [30] proposed a fully convolutional network (FCN) that extended DCNN image classification to dense pixel-wise classification by convolutionalization.

The main limitation of the FCN-based methods is low-resolution prediction due to multiple pooling operations. To resolve the limitation, there have been various approaches. One approach [42, 4] employed astrous convolution, also known as dilated convolution, which supports exponential expansion of the receptive field without loss of resolution. Chen *et al*. [4] additionally applied dense CRF method [21] to achieve detailed final prediction. Several follow-up studies [45, 29, 1, 3, 28] proposed sophisticated methods to combine CRF framework into DCNNs. Another approach [31, 2, 19] learned multiple deconvolution layers from low-resolution features to upsample the coarse feature map while recovering detailed boundaries.

The other approach [30, 12, 2, 33, 17, 26] exploited middle layer features to achieve high-resolution prediction. Long *et al*. [30] designed a skip architecture and merged score maps computed from multi-level features to obtain the final prediction. Hariharan *et al*. [12] constructed a feature vector called hypercolumn for every location by stacking features from some or all of the layers in the network. Several methods [2, 33, 17] applied skip-connections in feature upsampling procedures using deconvolution. In particular, Lin *et al*. [26] very recently achieved large improvement by designing a network called RefineNet that iteratively refines higher-level features by employing low-level features through residual connections. The network effectively conveys the low-level features as well as semantic high-level features and it can be efficiently trained end-to-end. Our RGB-D network revises this state-of-the-art architecture and takes the same advantage.

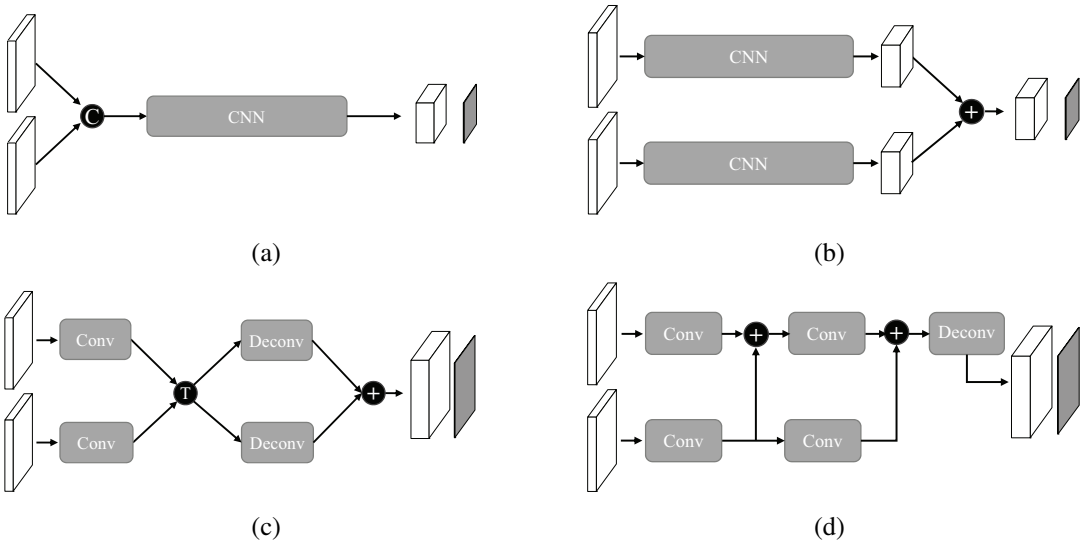

Figure 2. Different existing architectures for RGB-D semantic segmentation. (a) early fusion, (b) late fusion, (c) the architecture proposed by [39], (d) the architecture proposed by [13], where 'C', 'T', and '+' represent the concatenation, transformation, and element-wise summation, respectively.

For indoor semantic segmentation, a variety of methods utilizing depth information have been studied. Previously, most methods [32, 10, 20, 35, 6] computed hand-crafted features specifically designed for capturing depth features as well as color features. Then, they constructed a model to classify each region such as superpixel based on the features.

In contrast, recent methods [5, 11, 7, 25, 39, 13] usually employ DCNN that automatically trains features capturing different levels of representations. Couprie *et al*. [5] extended multi-scale RGB CNN architecture [8] to RGB-D situation by simply concatenating input color and depth channels, i.e., early fusion (Figure 2 (a)). Long *et al*. [30] additionally reported the result of fusing two predictions made by each RGB and depth modality, i.e., late fusion (Figure 2 (b)), as well as the result of early fusion. Gupta *et al*. [11] generalized the R-CNN system introduced by Girshick *et al*. [9] to leverage depth information. For that purpose, they encoded the depth image with three channels called HHA at each pixel: horizontal disparity, height above ground, and angle with gravity. Li *et al*. [25] captured and fused contextual information from RGB and depth features through bi-directional vertical and horizontal LSTM layers [38]. They used rather simple architecture especially for depth feature and partly utilized only RGB intermediate features through simple feature concatenation.

There have been encoder-decoder architectures [39, 13] similarly to RGB deconvolution-based methods. Wang *et al*. [39] proposed a structure for deconvolution of multiple modalities (Figure 2 (c)). It contains additional feature transformation network that correlates the two modalities by discovering common and modality specific features. It does not exploit any informative intermediate features of both modalities and it adopts simple score fusion of two modalities at the end of the network for final prediction. The training procedures consist of two stages rather than end-to-end. Hazirbas *et al*. [13] proposed a method that exploits intermediate depth features (Figure 2 (d)). However, as they simply sum intermediate RGB and depth features only in encoder part, it does not fully exploit effective mid-level RGB-D features, reporting accuracy worse than the state-of-the-art RGB-only CNN architecture [27].

In this paper, we propose a network that effectively exploits multi-level RGB and depth features simultaneously. Our network is trained to obtain optimal fusion of two complementary modality features through residual learning with skip-connection and iteratively refines the fused features. The multi-path residual feature fusion with skip-connection allows the backward gradient to easily propagate to both RGB and depth layers. In this way, the network trains end-to-end the discriminative RGB-D features which should be fused from low to high level.

3. Multi-level Residual Feature Fusion

Utilizing multi-level features is important for high resolution dense prediction. Existing RGB-D semantic segmentation approaches do not effectively extract or fuse those features in the two modalities. We propose a network that exploits multi-level RGB-D features and effectively fuses the features in different modalities through residual learning with skip-connections.

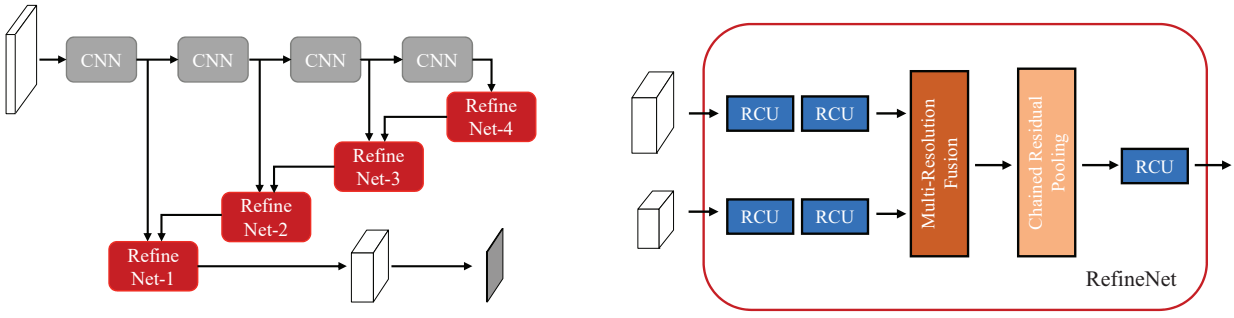

Figure 3. Building blocks of the network proposed by [26]. Left: network architecture for semantic segmentation. Right: detailed diagram of RefineNet block.

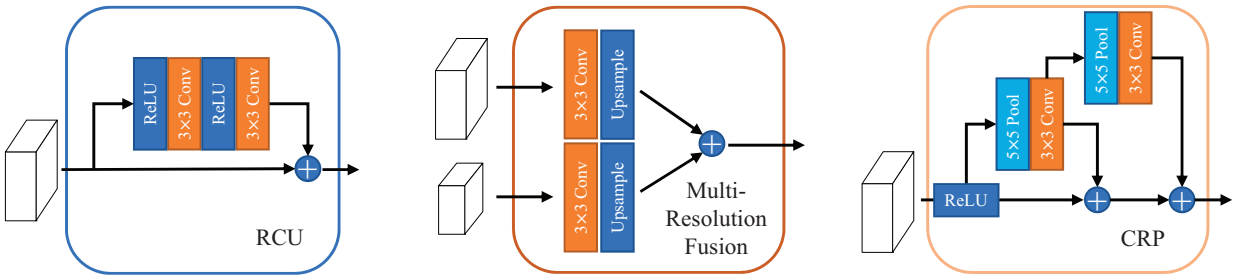

Figure 4. Details of the sub-modules in RefineNet.

In this section, we first review the recently proposed RefineNet architecture [26] that achieved great success in RGB semantic segmentation by employing residual connections. Then, we describe our network that extends the RefineNet to effectively train the way to extract and fuse multi-level RGB and depth features for indoor semantic segmentation.

3.1. Review of RefineNet

Recently ResNet [14, 15] has shown outstanding performance on image recognition. The simplest way to employ the ResNet to semantic segmentation is replacing the single label prediction layer with a dense prediction layer. However, it outputs prediction with 32 times smaller in each spatial dimension than the original image. To address the limitation, RefineNet iteratively refines higher-level features by incorporating low-level features through sub-building blocks, called RefineNet (Figure 3).

The RefineNet takes as inputs each multi-level ResNet feature through skip connection and the previously refined feature. Then, those features are refined and fused through a series of sub-components: residual convolutional unit, multi-resolution fusion, and chained residual pooling (Figure 4); The residual convolution unit (RCU) is an adaptive convolution set that fine-tunes the pretrained ResNet weights for semantic segmentation. The multi-resolution fusion block fuses the multi-path input into a higher-resolution feature map. One convolution in the block is for input adaptation, which matches the number of feature channels and re-scales the feature values appropriately for summation. The purpose of chained residual pooling (CRP) is to encode contextual information from a large region. The block consists of a chain of multiple pooling blocks, each consisting of one max-pooling layer and one convolution layer. The pooling operation has an effect that spreads the large activation values which can be accessed from nearby locations as contextual features. The additional convolution layer learns the importance of the pooled feature, which is fused to the original feature through residual connection. There is an additional RCU at the end of the RefineNet to employ non-linearity operations on the fused feature maps.

The core design philosophy of the RefineNet is motivated by the advantage of identity mapping with skip-connection [15]. The residual connections enable efficient backward propagation of gradients through RefineNet and facilitates end-to-end training of the multi-path network.

3.2. Our RDFNet with Multi-Modal Feature Fusion

The main issue of RGB-D semantic segmentation is how to effectively extract depth features along with color features and to utilize those features for the desired task of semantic segmentation. The RefineNet described in Section 3.1 proposed a generic means for fusing different levels of

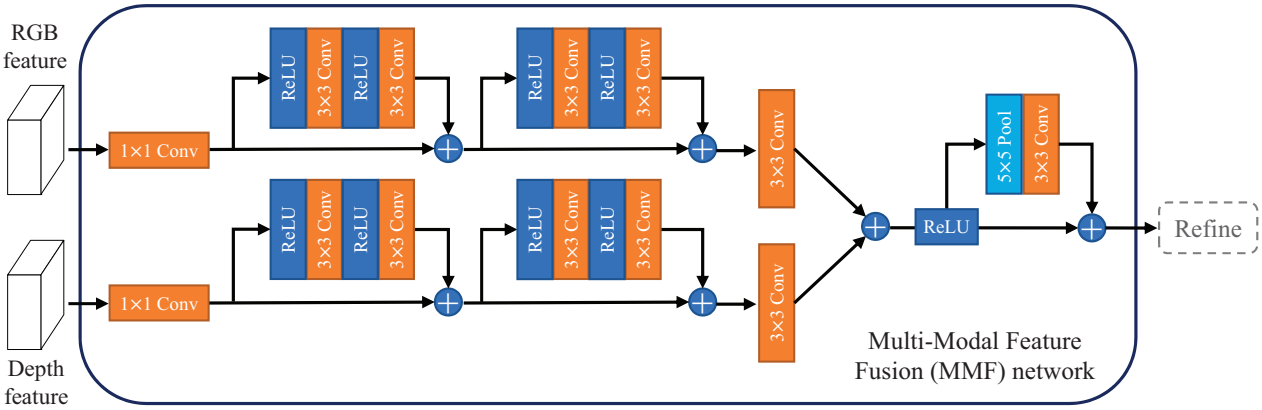

Figure 5. Diagram of our multi-modal feature fusion (MMF) network.

features, which is more effective than simple feature concatenation. In this paper, we employ a similar architecture for multi-modal CNN feature fusion while retaining the advantage of skip connection.

Our RDFNet extends the RefineNet to handle multi-modal feature fusion and includes RefineNet blocks for fused feature refinement. The overall diagram of our network is illustrated in Figure 1. Differently from existing networks that utilize depth information (Figure 2), our network is designed to fully exploit multi-level depth features through MMF blocks with an additional deep depth feature path based on ResNet [14].

The detailed components of our MMFNet is shown in Figure 5. Our feature fusion block consists of the same components as in RefineNet but with different inputs, from which desired operations are slightly different. Given RGB and depth ResNet features, our MMFNet first reduces the dimension of each feature through one convolution to facilitate efficient training while mitigating explosion of parameters. Then, each feature goes through two RCUs and one convolution as in RefineNet. There is a certain difference between the purpose of RCUs in MMFNet and those in RefineNet. The RCUs in our MMFNet are desired to perform some nonlinear transformations specifically for modality fusion. Two features in different modalities are complementarily combined to improve each other through the operations, where as those in RefineNet are mainly to refine coarse higher level feature by employing lower level feature with higher-resolution. Subsequent additional convolution in MMFNet is crucial to adaptively fuse features in different modalities as well as re-scaling the feature values appropriately for summation. As color features generally have better discrimination power than depth features for semantic segmentation, the summation fusion in the block mainly works to learn supplementary or residual depth features which might improve RGB features to discriminate confusing patterns. The importance of each modality feature can be controlled by the learnable parameters in the convolution after RCUs.

Finally, we perform an additional residual pooling operation to incorporate certain contextual information in the fused feature. We found one residual pooling in MMFNet of each level is enough. The stronger contextual information can be further incorporated in the following multi-level fusion through RefineNet blocks. Note that we skip the additional RCU at the end of original RefineNet in our MMFNet because the output of our MMFNet directly goes through the RCUs in the fore part of the RefineNet.

Our network is constructed to retain the philosophy of the RefineNet by employing residual learning with skip-connections through all the layers, which facilitates both of effective multi-level RGB and depth feature extraction and efficient end-to-end training.

3.3. Architecture details

Following the success of Gupta *et al*. [11], We encode the depthmap to a 3D image called HHA [10], which can be directly used as an input of the pre-trained network path for depth feature along with fine-tuning. The HHA representation encodes the properties of geocentric poses that emphasize complementary discontinuities in the image, which is hard to be trained through convolutional network. We compute depth features through ResNet with the same number of layers as RGB.

As depicted in Figure 1, we utilize 4-level RGB and depth features with different resolutions similarly to the RefineNet. We take *res5*, *res4*, *res3*, and *res2* features in ResNet [14] as inputs to our MMFNet. For each MMFNet, we include a dropout layer for regularization with ratio of 0.5 before 1×1 convolution. The MMFNet consists of ReLU nonlinearity, 3×3 convolution, and 5×5 pooling layer with stride of 1 and the number of filters (channels) in

the block is set to 512 for MMFNet-4 and 256 for the others. RefineNet blocks take the fused features and the previously refined feature as inputs except RefineNet-4 which only takes a fused feature from *res5*. The RefineNet-4 does not perform multi-resolution fusion. The number of filters in each RefineNet is set to the same as those of each MMFNet output. Final feature map obtained by RefineNet-1 goes through two additional RCUs, then another 1×1 convolution for the prediction with a dropout layer with ratio of 0.5. We add a softmax loss layer for loss computation. Our network with MMF blocks can be efficiently trained on a single GPU while fully utilizing the potentials of extremely deep RGB-D network.

4. Experiments

In this section, we evaluate our network through comprehensive experiments. We use two publicly available RGB-D datasets: NYUDv2 [35] and SUN RGB-D [37]. For the evaluation, we report three types of metrics (pixel accuracy, mean accuracy, and mean intersection over union (IoU)) widely-used to measure the performance of semantic segmentation [30]. As mentioned before, we use HHA encoding computed from a depthmap as our depth modality input.

4.1. Training details

We implemented our network using the publicly available Caffe toolbox [18] with an Nvidia GTX Titan X GPU. We employed general data augmentation schemes: random scaling, random cropping, and random flipping. We applied test-time multi-scale evaluation for all experiments by averaging the resulting predictions. We set the momentum and weight decay to 0.9 and 0.0005, respectively. We used the initial learning rate of 10^{-4} and divided it by 10 when the loss converges to a certain range and stops decreasing. We multiplied the learning rate by 0.1 for the base ResNet layers. All the parameters not in the base ResNet are initialized by a normal distribution with zero mean and 10^{-2} variance, while the biases were initialized with zero.

4.2. NYUDv2

NYUDv2 [35] is one of the most popular RGB-D dataset, which contains 1449 densely labeled pairs of RGB and depth images captured by using Microsoft Kinect. The dataset also provides inpainted depthmaps computed by the colorization method of Levin *et al*. [24], and we used the inpainted depthmaps for experiments. Following the standard train/test split, we use 795 training images and 654 test images. We evaluate our network for 40 classes using the labels provided by [10].

We first compare our RDFNet with the existing indoor semantic segmentation methods using CNN features. The results are shown in Table 1. It shows that our network outperforms all existing RGB-D methods as well as RGB methods, demonstrating that our network effectively utilizes depth information. It improves the accuracy of RGB-only RefineNet by 2.4%, 3.9%, and 3.6% for pixel accuracy, mean accuracy, and mean IoU, respectively.

	data	pixel acc.	mean acc.	IoU
Gupta *et al*. [11]	RGB-D	-	35.1	-
Eigen *et al*. [7]	RGB-D	65.6	45.1	34.1
FCN [30]	RGB-D	65.4	46.1	34.0
Wang *et al*. [39]	RGB-D	-	47.3	-
Context [27]	RGB	70.0	53.6	40.6
Refine-101 [26]	RGB	72.8	57.8	44.9
Refine-152 [26]	RGB	73.6	58.9	46.5
RDF-152 (ours)	RGB-D	**76.0**	**62.8**	**50.1**

Table 1. Semantic segmentation accuracy on NYUDv2. Our RDFNet outperforms all existing methods.

	pixel acc.	mean acc.	IoU
RDF-50	74.8	60.4	47.7
RDF-101	75.6	62.2	49.1
RDF-152	**76.0**	**62.8**	**50.1**

Table 2. Semantic segmentation accuracy on NYUDv2 of our network with variants of the pre-trained residual network.

As the multi-level features of our network are not limited to a specific pre-trained network, we report the accuracies of our network using residual networks with different number of layers, i.e., Res-50, Res-101, and Res-152. The results are shown in Table 2. It shows that the deeper the network becomes, the better results we generally get, while the amount of improvement decreases. It is noteworthy that the accuracy of our network with Res-50 using RGB-D data (RDF-50) is higher than those of RefineNet with Res-152 using RGB data (Refine-152 [26]).

Class-wise accuracies of our results compared with those of RefineNet are shown in Table 3. Our results show significant improvement in most categories by effectively employing depth features, especially in categories with clear geometric distinction such as *table*, *counter*, and *dresser*. The lower accuracy reported for the *board* class is due to the fact that there are few images containing boards in the dataset. It is also hard to improve the discrimination between a board and a picture with little geometric differences even using additional depth features.

We validate our network in Table 4 by comparing with other variants. Here we use Res-101 for the experiments. We first report the accuracies of depth-only networks to show that the RefineNet also works properly for extracting depth features from HHA encoding, which validates our choice of depth feature part. We trained a RefineNet model based on ResNet features finetuned using only HHA input. The accuracy of RefineNet only using HHA (Refine-HHAonly) is even higher than FCN using both RGB and HHA. This result demonstrates that ResNet with finetuning

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bkshelf
Refine-101 [26]	77.5	82.9	58.7	65.7	59.1	57.8	40.1	36.7	45.8	42.8
RDF-101	78.8	**87.3**	**63.0**	71.6	**65.1**	62.8	49.7	39.5	48.5	**46.5**
RDF-152	**79.7**	87.0	60.9	**73.4**	64.6	**65.4**	**50.7**	**39.9**	**49.6**	44.9
	picture	counter	blind	desk	shelf	curtain	dresser	pillow	mirror	mat
Refine-101 [26]	60.1	56.8	61.4	22.6	12.3	53.5	38.3	39.6	38.7	29.7
RDF-101	60.8	65.5	61.5	**30.8**	12.4	54.0	**54.0**	46.6	**55.5**	**41.6**
RDF-152	**61.2**	**67.1**	**63.9**	28.6	**14.2**	**59.7**	49.0	**49.9**	54.3	39.4
	cloths	ceiling	books	refridg	tv	paper	towel	shower	box	board
Refine-101 [26]	24.4	66.0	33.0	52.4	52.6	31.3	36.8	23.6	11.1	**63.7**
RDF-101	26.3	**69.7**	**36.0**	55.7	63.2	**34.6**	39.1	**38.5**	**13.1**	46.0
RDF-152	**26.9**	69.1	35.0	**58.9**	**63.8**	34.1	**41.6**	**38.5**	11.6	54.0
	person	stand	toilet	sink	lamp	bathtub	bag	othstr	othfurn	othprop
Refine-101 [26]	78.6	38.6	68.4	53.2	45.9	32.9	14.6	**32.9**	18.7	36.4
RDF-101	**81.8**	42.5	**68.9**	56.1	45.8	49.0	13.4	31.0	19.5	38.6
RDF-152	80.0	**45.3**	65.7	**62.1**	**47.1**	**57.3**	**19.1**	30.7	**20.6**	**39.0**

Table 3. Class-wise semantic segmentation accuracy (IoU) on NYUDv2.

	pixel acc.	mean acc.	IoU
FCN32-HHAonly	58.3	35.7	25.2
Refine-HHAonly	66.5	46.5	36.3
Refine-Concat	74.5	59.2	47.0
RDF-101	**75.6**	**62.2**	**49.1**
-Without RP	75.4	61.1	48.7
-Without conv	74.7	59.6	47.7
-Without skip	73.8	58.7	45.8
RDF-101-depth	75.3	60.9	48.2

Table 4. Comparison for different variants of network.

	data	pixel acc.	mean acc.	IoU
Ren *et al.* [32]	RGB-D	-	36.3	-
B-SegNet [19]	RGB	71.2	45.9	30.7
LSTM [25]	RGB-D	-	48.1	-
FuseNet [13]	RGB-D	76.3	48.3	37.3
Context [28]	RGB	78.4	53.4	42.3
Refine-152 [26]	RGB	80.6	58.5	45.9
RDF-152 (ours)	RGB-D	**81.5**	**60.1**	**47.7**

Table 5. Semantic segmentation accuracy on SUN RGB-D. Our RDFNet achieves the state-of-the-art accuracy.

can extract appropriate features from depth data.

We also compare our MMFNet with a baseline fusion method. For the comparison, we replace our MMFNet with feature concatenation fusion with additional dropout layer and one convolution layer for dimension reduction. Here we only compare with the multi-level concatenation fusion (Refine-Concat) because we found that it generally shows better accuracy than other fusion architectures (early fusion, late fusion, and other variations). Note that the results show that our MMFNet effectively utilizes multi-modal features, achieving higher accuracies for all metrics, specifically by 1.1%, 3.0%, and 2.1%, respectively. It confirms that the improvement specifically comes from MMF rather than simple addition of depth information.

We additionally conduct ablative experiments for our MMFNet by successively eliminating each component (Table 4). Without residual pooling (Without RP) the accuracy decreases slightly, which means the additional residual pooling is rather optional. We found further pooling did not improve the accuracy. However, the experiments show that the other components are crucial for effective feature fusion. Without the convolution (Without conv) that adaptively controls the weight to fuse each modality feature, we obtained much less accuracy while it is only slightly higher than those of concat fusion. We additionally report the accuracy without skip connection in RCUs (Without skip). Here the features directly go through the nonlinearity transformations and sum fusion. By comparing the accuracies, we can see the importance of skip connection for effective end-to-end training of multi-level features.

We finally report the result of our network trained directly on depth data instead of HHA to show that our network can be applied to different types of RGBD inputs. We preprocessed the depth to roughly scale the values into the range of $0 \sim 255$. Specifically, we simply used $k/depth$, similarly to the disparity channel in HHA, where k is a constant. The result (RDF-101-depth) shows consistent improvement over RefineNet while slightly worse than our RDFNet with HHA (RDF-101). It indicates that our RDFNet can efficiently learn to extract meaningful features directly from the depth data as well.

4.3. SUN RGB-D

SUN RGB-D dataset [37] has been built for a large-scale RGB-D benchmark. The dataset consists of 10335 pairs of

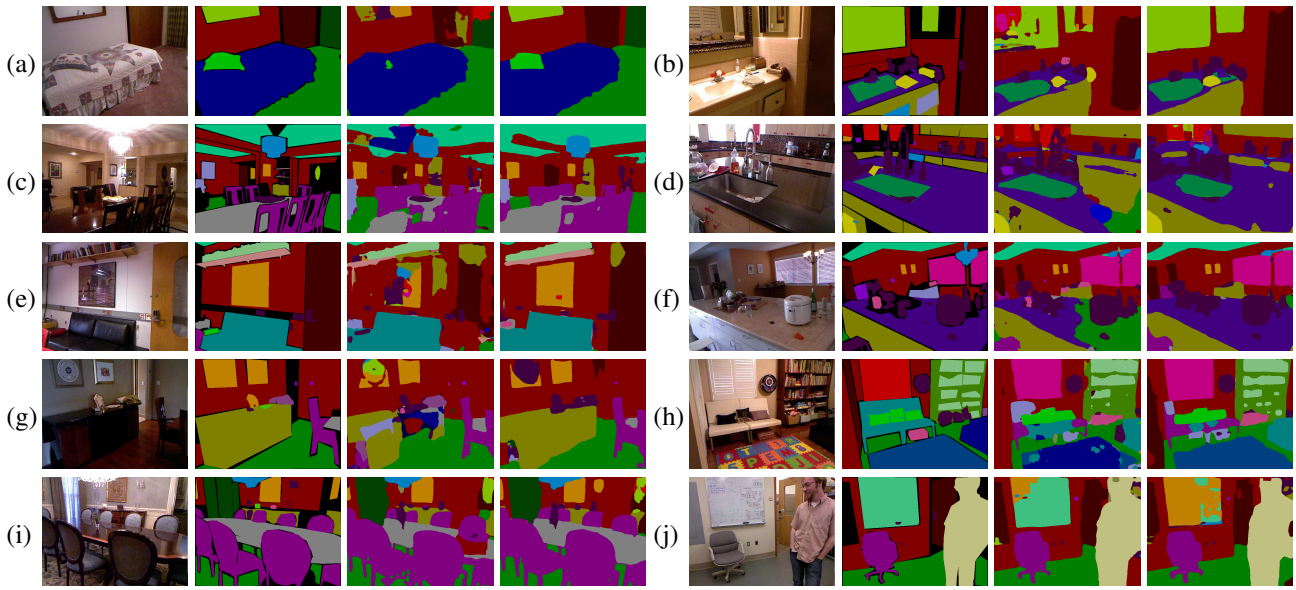

Figure 6. Qualitative results of our RDFNet compared with RefineNet [26]. From left to right for each example: image, ground truth, the results obtained by RefineNet, and ours. Note that the depth features help to discriminate regions that might be confusing only with color features, e.g., pillow with patterns similar to bed (a), the door with a homogeneous pattern (b, e), ceiling with clear geometric distinction (c), counter with vertical surface normal (d, f), cabinet with low illumination (g), mirror reflecting other color patterns (a, b), floor mat on the floor (h), and top surface of a table (c,i). The last example shows a failure case of ours (j). Best viewed in color.

RGB and depth images captured from four different depth sensors, which contains images from NYUDv2 depth [35], Berkeley B3DO [16], and SUN3D [40], as well as newly captured images. We use the standard split of 5285 training images and 5050 test images with pixel-wise labeling of 37 classes for evaluation.

Table 5 shows that our network outperforms existing RGB-D methods by a large margin. It also achieves the state-of-the-art accuracy for all metrics, improving the accuracy of RGB-only RefineNet by a considerable amount. The ability of depth feature might be slightly diminished for this dataset because it contains many bad depth images with invalid measurements, e.g., images obtained by RealSense RGB-D camera. Nevertheless, the results demonstrate that our network learns effective RGB-D features on a large-scale dataset even without manually weeding the bad images out.

4.4. Qualitative results

We show some qualitative results of ours compared with RefineNet [26] in Figure 6. We obtained the results of the RefineNet by running the publicly available source code with the provided model based on Res-101. We compare the results with our RDF-101 using RGB-D inputs. The comparisons illustrate that our network effectively utilizes depth features to discriminate regions that might be confusing with only color features.

5. Conclusion

We proposed a novel network that takes full advantage of residual learning with skip-connection to extract effective multi-modal CNN features for semantic segmentation. The residual architecture facilitates efficient end-to-end training of very deep RGB-D CNN features on a single GPU. Our MMFNet shows that the recent multi-level feature refinement architecture [26] can be effectively extended to utilize features in different modalities, while retaining the advantage of skip-connection. Our experiments demonstrated that the proposed network outperforms existing methods, obtaining the state-of-the-art mean IoUs of 50.1% and 47.7% for NYUDv2 and SUN RGB-D indoor datasets, respectively.

Acknowledgements This work was supported by the Ministry of Science and ICT, Korea, through IITP grant (R0126-17-1078), Giga Korea grant (GK17P0300), and NRF grant (NRF-2014R1A2A1A11052779).

References

[1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr. Higher order conditional random fields in deep neural networks. In *Proc. ECCV*, pages 524–540. Springer, 2016. 1, 2

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 2

[3] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *Proc. ECCV*, pages 402–418. Springer, 2016. 1, 2

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1, 2

[5] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 1, 2, 3

[6] Z. Deng, S. Todorovic, and L. Jan Latecki. Semantic segmentation of rgbd images with mutex constraints. In *Proc. ICCV*, pages 1733–1741, 2015. 1, 2, 3

[7] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV*, pages 2650–2658, 2015. 1, 2, 3, 6

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013. 3

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, pages 580–587, 2014. 3

[10] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proc. CVPR*, pages 564–571, 2013. 1, 2, 3, 5, 6

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proc. ECCV*, pages 345–360. Springer, 2014. 1, 2, 3, 5, 6

[12] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proc. CVPR*, pages 447–456, 2015. 2

[13] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Proc. ACCV*, volume 2, 2016. 1, 2, 3, 7

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1, 2, 4, 5

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Proc. ECCV*, pages 630–645. Springer, 2016. 2, 4

[16] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013. 8

[17] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *arXiv preprint arXiv:1611.09326*, 2016. 2

[18] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[19] 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. 2, 7

[20] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Geometry driven semantic labeling of indoor scenes. In *Proc. ECCV*, pages 679–694. Springer, 2014. 1, 2, 3

[21] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. NIPS*, 2011. 2

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012. 1, 2

[23] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *Proc. ECCV*, pages 424–437. Springer, 2010. 1

[24] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM, 2004. 6

[25] 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 *Proc. ECCV*, pages 541–557. Springer, 2016. 1, 2, 3, 7

[26] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017. 2, 4, 6, 7, 8

[27] G. Lin, C. Shen, A. v. d. Hengel, and I. Reid. Exploring context with deep structured models for semantic segmentation. *arXiv preprint arXiv:1603.03183*, 2016. 3, 6

[28] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. CVPR*, pages 3194–3203, 2016. 1, 2, 7

[29] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proc. ICCV*, pages 1377–1385, 2015. 1, 2

[30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. CVPR*, pages 3431–3440, 2015. 1, 2, 3, 6

[31] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. ICCV*, pages 1520–1528, 2015. 2

[32] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Proc. CVPR*, pages 2759–2766. IEEE, 2012. 1, 2, 3, 7

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2

[34] C. Russell, P. Kohli, P. H. Torr, et al. Associative hierarchical crfs for object class image segmentation. In *Proc. ICCV*, pages 739–746. IEEE, 2009. 1

[35] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. ECCV*, pages 746–760. Springer, 2012. 1, 2, 3, 6, 8

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[37] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. CVPR*, pages 567–576, 2015. 6, 7

[38] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015. 3

[39] 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 *Proc. ECCV*, pages 664–679. Springer, 2016. 1, 2, 3, 6

[40] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proc. ICCV*, pages 1625–1632, 2013. 8

[41] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Proc. CVPR*, pages 702–709. IEEE, 2012. 1

[42] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1, 2

[43] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, pages 818–833. Springer, 2014. 1, 2

[44] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1

[45] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proc. ICCV*, pages 1529–1537, 2015. 1, 2