Variational Context-Deformable ConvNets for Indoor Scene Parsing

Zhitong Xiong, Yuan Yuan*, Nianhui Guo, Qi Wang
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, Shaanxi, P. R. China
{xiongzhitong, y.yuan1.ieee, guonianhui199512, crabwq}@gmail.com

Abstract

Context information is critical for image semantic segmentation. Especially in indoor scenes, the large variation of object scales makes spatial-context an important factor for improving the segmentation performance. Thus, in this paper, we propose a novel variational context-deformable (VCD) module to learn adaptive receptive-field in a structured fashion. Different from standard ConvNets, which share fixed-size spatial context for all pixels, the VCD module learns a deformable spatial-context with the guidance of depth information: depth information provides clues for identifying real local neighborhoods. Specifically, adaptive Gaussian kernels are learned with the guidance of multimodal information. By multiplying the learned Gaussian kernel with standard convolution filters, the VCD module can aggregate flexible spatial context for each pixel during convolution. The main contributions of this work are as follows: 1) a novel VCD module is proposed, which exploits learnable Gaussian kernels to enable feature learning with structured adaptive-context; 2) variational Bayesian probabilistic modeling is introduced for the training of VCD module, which can make it continuous and more stable; 3) a perspective-aware guidance module is designed to take advantage of multi-modal information for RGB-D segmentation. We evaluate the proposed approach on three widely-used datasets, and the performance improvement has shown the effectiveness of the proposed method.

1. Introduction

Pixel-wise semantic image analysis, e.g. semantic segmentation, is a fundamental computer vision task with numerous applications, such as robot sensing and autonomous driving. Compared with image classification, semantic segmentation needs to distinguish the category of each pixel under complicated backgrounds. As semantic segmentation is to understand the image at pixel-level, high semantic-level features are critical for the pixel-wise classification task. Thus, deep Convolutional Neural Networks(CNNs) have significantly boosted the performance of semantic segmentation. Fully Convolutional Networks (FCNs) [39] based methods and Encoder-Decoder architecture based methods achieve remarkable results on several public benchmarks. Although the segmentation performance can be improved significantly by employing deep features, there are still challenges for indoor scene semantic segmentation: 1) more complex spatial layout and cluttered objects make the indoor scene segmentation challenging; 2) fixed receptive-field may cause feature-inconsistency between different objects; 3) using 2D image alone for image understanding may cause geometric confusions. Fortunately, depth sensor can provide geometric information to aid the pixel-wise understanding of indoor scenes.

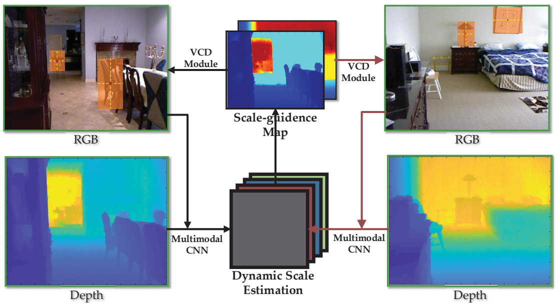

Figure 1. Illustration of context-deformable convolution. The scale-guidance maps are learned with the guidance of multi-modal features.

Commercial RGB-D sensors such as Miscrosoft Kinect [53] become more and more popular. Adapting conventional 2D image analysis tasks such as object detection [42], semantic segmentation to RGB-D image becomes feasible. Since RGB-D image can provide illumination-insensitive geometric cues, extensive methods [15], [40], [35], and [27] have shown that enhanced results can be obtained by incorporating the information from depth modality. Depth information is useful for describing the 3D scene layout,

*Corresponding Author.

and properly extracted depth features can represent the geometric information of the image. A variety of methods have been explored for fusing the geometric feature with RGB appearance feature together for RGB-D segmentation. Image-level [7] and deep feature-level fusion [16, 29, 21] were designed to enhance the multi-modal feature fusion. However, merely fusing depth and RGB features cannot fully exploit the geometric information provided by depth modality. Since depth information is helpful for identifying the real physical scales of objects, depth modality can be used to learn scale-adaptive features by adjusting the receptive-field of each pixel.

Towards perspective understanding of RGB-D scene, Depth-aware CNNs [47] built a depth-aware receptive field to augment standard convolution. However, exploiting local depth similarity lacks global understanding of the scene layout for learning explicitly scale-adaptive context. Cascaded Feature Network [30] split the RGB image into different depth-resolution for a fine-grained semantic segmentation. Depth-aware gating module [25] learned features of dynamic-scale by gating on multi-scale feature pyramids. However, using multi-branch architecture is time-consuming and not flexible.

We aim to take advantage of depth modality to learn structured deformable receptive-fields for convolution. As shown in Fig. 1, depth and RGB modality are both exploited to predict a location-variant scale-guidance map for context-deformable convolution. Specifically, we propose a variational context-deformable (VCD) convolution module, which augments standard convolution by a structured learnable spatial Gaussian kernel. The spatial Gaussian kernel is learned with the guidance of both RGB and depth modality, which can adjust receptive-field by multiplying a Gaussian mask on neighboring pixels. The VCD module allows spatially varying convolution over the fixed receptive-field by multiply with a variance-learnable Gaussian kernel. By adjusting the variance of Gaussian kernel, the VCD module can capture proper context for different pixels.

Additionally, to make the proposed context-deformable module more stable and continuous, we enhance this module with Bayesian probabilist modeling, i.e., making the learned scale-guidance map (variances for generating Gaussian kernel weights) a random variable rather than deterministic values. The proposed method has the following advantages:

1. The receptive-field can be learned and adjusted in a more compact and structured way, which can be easily optimized. Moreover, controlling the receptive-field with learnable Gaussian mask is more interpretable.

2. Modeling the variance for generating Gaussian mask as latent random variables can make the training of VCD module more stable and enforce the learned receptive-field to be continuous. By stacking multiple VCD layers, the final receptive-field for each pixel can be continuously adjusted.

3. Multi-modal information provides important cues for generating the scale-guidance map, which is an effective and flexible way for exploiting the strong depth prior information.

4. The proposed VCD module can be readily plugged in any deep architectures to replace standard convolutions efficiently for semantic segmentation.

2. Related Work

2.1. RGB-D Image Semantic Segmentation

RGB-D indoor scene parsing has been studied for years, and numerous methods have been proposed [7, 13, 45, 16, 31]. At the early stage, hand-crafted depth feature extraction methods were designed [14, 46, 45]. Gupta et al. [16] proposed to encode the depth image with three channels as HHA image, which was widely applied to many RGB-D analysis tasks. Multi-modal fusion based methods are the most popular approaches for exploiting depth information, which can be conducted at image-level or feature level [21]. FuseNet [17] summed RGB and depth features to obtain multi-modal representations. Multi-level feature fusion was proposed by [37] to extend the residual learning idea of RefineNet [32] for RGB-D semantic segmentation. ACNet [19] and Gated Fusion[4] were designed to automatically learn the contributions of each modality for scene segmentation in different scenes. Mutex Constraints was proposed by [9] to make use of 3D geometric structure for RGB-D segmentation. Geometry-Aware Distillation [22] proposed to jointly infer the semantic and depth information by distilling geometry-aware embedding.

2.2. Adaptive Context for Segmentation

Depth-aware CNN [47] exploited depth-similarity to augment standard convolution with depth-related local context. Pixel-adaptive convolution [43] enhanced the vanilla convolution by multiplying with a spatially varying kernel that depends on local pixel features. SurfConv [5] was proposed to exploit 3D information by a depth-aware multi-scale 2D convolution, which was proven to be effective for 3D semantic segmentation task. 3D Neighborhood Convolution [3] introduced an effective way to model the receptive field of 2D convolution based on the locality from the 3D real world space. Scale-adaptive convolution [52] designed a scale regression layer, which resized the convolutional patches adaptively to tackle the feature-inconsistent problem. Cascaded Feature Network [30] split the RGB image into several layers according to depth image and exploited multi-branch architecture to get a fine-grained semantic segmentation. Deformable convolution networks (DCN) [8] was introduced to augment convolution with learnable offsets. Deformable ConvNets v2 [56] reformulated DCN with mask weights, which alleviated the influence of irrelevant image content. OCNet [51] was proposed to exploit pixel features that belonged to the same object category as the object context for semantic segmentation. Semantic-correlation context [10] shared a similar idea to [51], which employed paired convolution to infer the semantic-correlated shape-context for segmentation. Non-local network based methods [57, 20, 12] aimed to aggregate features of all positions with a learned weight, which also exploited dynamic context features.

3. Methodology

3.1. Context-deformable Convolution

The standard convolution operation shares a fixed receptive-field for each pixel across the whole image. Aggregating context information with regular sliding window has an obvious limitation: the representation for objects with different scales will be inconsistent. This makes standard convolution less effective for handling the scale-variant problem. First of all, we will introduce the widely-used standard convolution. Given the input feature map $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$ and kernel weight $\mathbf{W} \in \mathbb{R}^{c_o \times c \times k \times k}$, the standard convolution can be formulated as:

$$\mathbf{v}_i = \sum_{\mathbf{p}_j \in \Omega(i)} \mathbf{W}[\mathbf{p}_j - \mathbf{p}_i] \cdot \mathbf{x}[\mathbf{p}_j], \tag{1}$$

where $v_i \in \mathbb{R}^c$ is the output features at pixel i, and $[\mathbf{p}_j - \mathbf{p}_i]$ denotes indexes of the 2D spatial offsets. For pixel i, $k \times k$ surrounding pixels (denoted by $\mathbf{p}_j \in \Omega(i)$) are aggregated with weights W. Thus, a fixed local context with size $k \times k$ is used for every pixel, which limits the flexibility of convolution operation and makes it hard to handle feature-inconsistency problem between large and small objects.

Although several works such as pixel-adaptive convolution [43] and Shape-Variant Context methods[10] have been proposed for variant context modeling, these methods involve in computing the pixel-wise feature-similarity, which will increase the computation cost significantly when using them at multiple layers. Different from them, in this work, a structured context-deformable convolution is proposed, which augments vanilla convolution with Gaussian kernel mask to control the effective receptive-field softly. The proposed context-deformable convolution can be formulated as:

$$\mathbf{v}_i = \sum_{\mathbf{p}_j \in \Omega(i)} GK(\mathbf{p}_i, \mathbf{p}_j)\mathbf{W}[\mathbf{p}_j - \mathbf{p}_i] \cdot \mathbf{x}[\mathbf{p}_j], \tag{2}$$

where $GK(\mathbf{p}_i, \mathbf{p}_j; g_\sigma^i)$ is a learnable spatial Gaussian kernel with parameter g_σ^i as standard deviation at pixel i. The Gaussian distribution function $GK(\mathbf{p}_i, \mathbf{p}_j; g_\sigma^i)$ can be defined as follows:

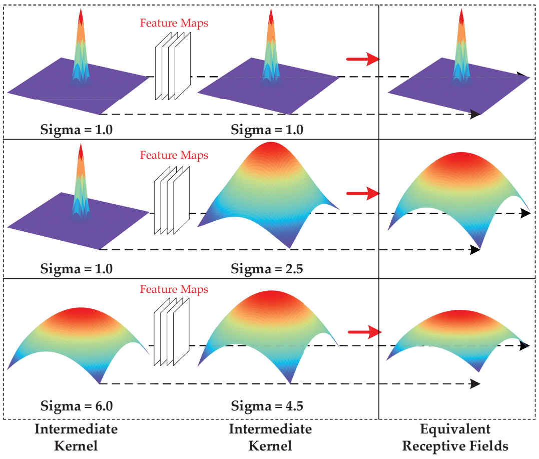

Figure 2. Stacking multiple different VCD modules can result in continuously adjusted receptive-field. Take the first row for example: multiple small contexts result in a relatively small receptive-field.

$$GK(\mathbf{p}_i, \mathbf{p}_j; g_\sigma^i) = \frac{1}{\sqrt{2\pi} g_\sigma^i} \exp(-\frac{(\mathbf{p}_j - \mathbf{p}_i)^2}{2{g_\sigma^i}^2}), \tag{3}$$

where g_σ^i is the i^{th} element of the scale-guidance map to control the scale of the spatial Gaussian kernel at pixel i. Specifically, for one convolution layer, $h \times w$ position-wise parameters is predicted densely, i.e., the scale-guidance map g_σ, which can be used to adjust the spatial context of each pixel. We expect that for a large context, the learned g_σ^i should be large to generate flat weights equivalent to standard convolution. While for a small context, the Gaussian kernel weights are expected to be uneven and concentrate on pixel i. Take, for example, a convolution layer with 1×3 kernel size: the learned Gaussian kernel weights for a large context may be $[0.99, 1.0, 0.99]$. For small context, the learned Gaussian weights may be $[0.2, 1.0, 0.2]$. In this way, the receptive-field for each pixel can be adjusted adaptively.

To disentangle the spatial context into scale deformation (VCD) and shape (DCN) deformation, we integrate the proposed VCD module with DCN [8]. Since DCN adjusts the spatial context by predicting K offsets, it is more suitable to learn the shape-deformation. VCD adjusts the scale context in a more compact and structured way, which provides constraints on the offset learning in DCN. This can be defined as:

$$\mathbf{v}_i = \sum_{\mathbf{p}_j \in \Omega(i)} GK(\mathbf{p}_i, \mathbf{p}_j)\mathbf{W}[\mathbf{p}_j - \mathbf{p}_i] \cdot \mathbf{x}(\mathbf{p}_j + \Delta\mathbf{p}_j), \tag{4}$$

where $\Delta\mathbf{p}_j$ is the learned 2D offsets, and $\mathbf{x}(\cdot)$ is the differential bilinear sampling operation. It is worth mentioning that using the relative position $\mathbf{p}_i$ and $\mathbf{p}_j$ can keep the spatial-structure of the initial Euclidean Space. In this way, the model is simplified and thus can make the training more stable.

Stacking multiple convolution layers can increase the receptive-field size linearly [34], and each convolution layer increases the receptive-field size by the kernel-size. Thus, adjusting the kernel size of each CNN layer adaptively can result in a more flexible and continuous deformable spatial-context. As illustrated in Fig. 2, stacking multiple context-deformable modules can result in a continuously spatial-variant context for one image pixel.

3.2. Variational Inference on Gaussian Variance

To adjust the receptive-field continuously, modeling the aforementioned context-deformable convolution in a Bayesian probabilistic framework is an effective way. Specifically, we model the parameter g_σ as a latent random variable, which enforces the module to learn distributions rather than deterministic values. The introduced regularization term can also enhance the correlation between spatial-context and segmentation performance. Thus, by modeling the parameter g_σ in a Bayesian probabilistic framework, the proposed method will be more stable, continuous and allowing interpolation.

Suppose the segmentation dataset with N samples is: $D = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $x^{(i)}$ and $y^{(i)}$ are the image and label respectively. Instead of using deterministic value, we aim to sample g_σ from a posterior distribution $p_\theta(g_\sigma|x)$ to generate the spatial Gaussian kernel using GK. However, the estimation of the true posterior density $p_\theta(g_\sigma|x)$ is intractable due to the computation-intractable integrals. Thus, the widely-used variational inference [24] is employed to approximate $p_\theta(g_\sigma|x)$ with a distribution $q_\phi(g_\sigma)$. As pointed by [24], maximizing the likelihoods of datapoints is equivalent to maximizing the variational evidence lower bound (ELBO) $\mathcal{L}(\theta, \phi; x^{(i)})$, which can be defined as:

$$\begin{aligned}\mathcal{L}(\theta, \phi; x^{(i)}) = \mathbb{E}_{q_\phi(g_\sigma|x^{(i)})}[\log p_\theta(y^{(i)}|x^{(i)}, g_\sigma)] \\ - KL(q_\phi(g_\sigma|x^{(i)})||p_\theta(g_\sigma)),\end{aligned} \tag{5}$$

where the first expected log-likelihood term is to minimize the segmentation loss. It can also be called prediction loss. The second Kullback-Leibler divergence term is used to minimize the distance of two distributions, i.e., $q_\phi(g_\sigma)$ and the true posterior $p_\theta(g_\sigma|x)$. This variational evidence lower bound can also be interpreted as minimizing the segmentation prediction loss, and simultaneously enforcing the variable g_σ to obey a defined distribution.

To optimize the segmentation loss term with the gradient descent method, we need to compute the gradient of it w.r.t. the parameter ϕ. In this work, Gaussian distribution is chosen as the approximate distribution: $q_\phi(g_\sigma) = \mathcal{N}(\mu, \sigma^2)$.

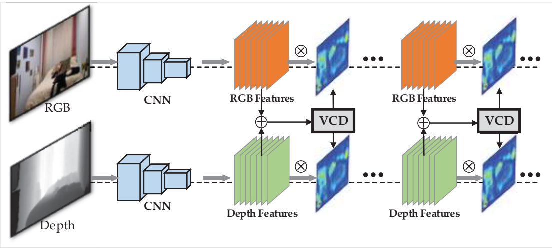

Figure 4. Illustration of the multi-modal guided scale-adaptive network architecture for RGB-D semantic segmentation.

Thus, the learnable parameter ϕ of the approximate distribution are μ and σ, which are predicted by two CNN layers in this work. Since the gradient cannot be computed directly, the reparameterization trick is employed to sample $g_\sigma \sim q_\phi(g_\sigma|x^{(i)})$ alternatively: $g_\sigma = \mu + \sigma\varepsilon$, where ε is the standard normal distribution, namely, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. By using the reparameterization trick, the ELBO can be rewritten as:

$$\begin{aligned}\mathcal{L}(\theta, \phi; x^{(i)}) = \frac{N}{M}\sum_{i=1}^{M}(\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(y^{(i)}|x^{(i)}, \mu + \sigma\varepsilon^l) \\ - KL(q_\phi(g_\sigma|x^{(i)})||p_\theta(g_\sigma))),\end{aligned} \tag{6}$$

where M is the number of mini-batch samples randomly drawn from N datapoints. L is the number of samples for one datapoint (image and label pair). In this way, the expected prediction loss term can be optimized with SGD optimizer.

For the prior distribution, the commonly used Gaussian distribution is chosen in this work, which can be defined as:

$$p_\theta(g_\sigma) = \mathcal{N}(\mu^*, \mathbf{I}), \tag{7}$$

where μ^* is set identical to μ in q_ϕ. Since we aim to model g_σ as a random variable to make the context-deformable convolution more stable and allow interpolation, the variance of ε is set to 1. As depicted above, we choose Gaussian distribution as the approximate distribution. Thus, the KL divergence loss can be formulated analytically as:

$$KL(q_\phi(g_\sigma|x^{(i)})||p_\theta(g_\sigma)) = \sum \log\frac{1}{\sigma} + \frac{\sigma^2 - 1}{2}. \tag{8}$$

With Eq. 6 and 8, the proposed VCD module can be trained in an end-to-end manner. The illustration of the VCD module is shown in Fig. 3. For each VCD layer, g_σ is first sampled from the learned distribution q_ϕ. Then it will be used to generate Gaussian kernel weights using $GK(p_i, p_j; g_\sigma)$. Finally, the Gaussian weights are multiplied with convolution weights in an element-wise manner for each pixel.

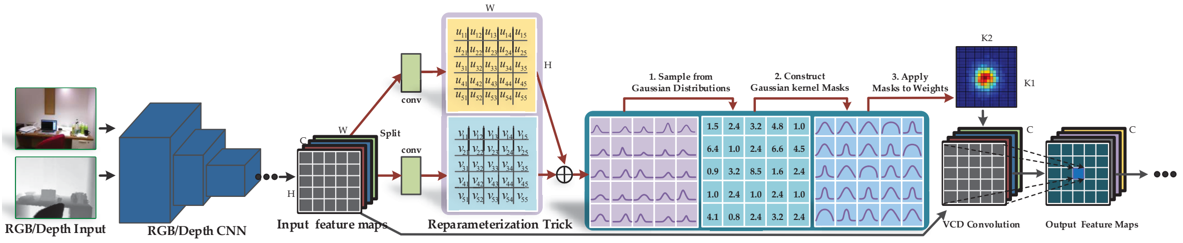

Figure 3. Illustration of the proposed VCD module. Two convolution layers are used for predicting μ and σ. Then scale-guidance map g_σ is sampled for generating the spatial Gaussian kernel weights. By multiplying the Gaussian weights with convolution filters, the spatial-context can be adjusted in a soft way.

3.3. Multi-modal Guided Scale-Adaptive Network

In this section, we mainly introduce the design of the probabilistic encoder $q_\phi(g_\sigma|x^i)$. Since convolutional neural networks are effective for pixel-wise feature learning, a CNN-based probabilistic encoder is introduced in this work. Although depth information has great correlation with the object-scale, merely using depth modality lacks the appearance information of objects. RGB modality can provide rich image content cues, which are also important for predicting the scale-guidance map g_σ densely. Considering this, multi-modal CNN features are exploited for learning the scale-guidance map in RGB-D semantic segmentation task.

As shown in Fig. 3, for one VCD layer, two extra convolutions are employed to output the distribution parameters μ and σ, which can be defined as:

$$\begin{aligned} \log \mu &= conv_\mu(F_{rgb} + F_d), \\ \sigma &= conv_\sigma(F_{rgb} + F_d), \end{aligned} \tag{9}$$

where $conv_\mu$ and $conv_\sigma$ are two convolutional layers with kernel size 3×3. F_{rgb} and F_d are the intermediate deep features of RGB and depth modality, respectively. The exponential function is used to guarantee that all the values in scale-guidance map g_σ are positive. We illustrate the VCD module based RGB-D segmentation framework in Fig. 4. Multiple VCD modules are stacked to form a deep network. The scale-guidance map for each module is predicted using the deep multi-modal features. Then RGB and depth modality share the same scale-guidance map for learning scale-deformable features.

Additionally, VCD-module can also be applied to RGB semantic segmentation. In this case, only RGB features are used for predicting the scale-guidance map. In practice, we choose to replace the standard convolution with VCD module at relatively deeper layers from two reasons: one is that deeper CNN features are with higher semantic-level, so they can be used to predict more accurate scale-guidance map based on the image content and depth information. The other reason is that convolution at deeper layers can affect larger receptive-field.

As semantic segmentation task needs high computation cost, a large L will make the whole network slow to converge. In this case, we use a two-stage optimization method for training the VCD module. At the first stage, all the parameters g_σ in scale-guidance map are set as deterministic values. Therefore, the KL divergence loss is not optimized, and L is 1 in this case. When the model converges, reasonable scale-guidance maps can be obtained. Then we treat the learned deterministic scale maps as the μ of distribution $q_\phi(g_\sigma)$ for the second stage training. At the second stage, the whole model can be optimized with Eq. 6 for training.

4. Experiments

To evaluate the proposed method comprehensively, experiments on two widely used public RGB-D semantic segmentation dataset: NYUv2 [36] and SUN RGB-D [41]. In addition to RGB-D indoor segmentation, we also evaluate VCD module on Cityscapes street scene dataset [6].

4.1. Datasets

As for RGB-D datasets, there are 1,449 RGB-D indoor image pairs with finely annotated segmentation labels in NYUv2 dataset. To fairly compare with previous methods, 795 images are split for training, and 654 images for testing. Following the experiment setting in [15], we use segmentation labels with 40 classes for all the experiments.

SUN RGB-D dataset is a large scale RGB-D dataset with 10,335 images. All of the images are annotated with 37 categories. Following the previous works, 5,285 RGB-D image pairs are split as training images, and the rest 5,050 images are used for testing.

Cityscapes benchmark is designed for urban segmentation. 5,000 high-resolution images (1024×2048) are finely annotated into 19 categories. Moreover, an extra 20,000 coarsely annotated images are also offered. All these images are captured from 50 different cities. In this work, only the finely annotated images are used, including 2,975 images for training and 500 images for validation. The rest 1,525 images without labels are used for testing on the benchmark server. To evaluate the performance of semantic

Table 1. Performance Comparisons on Sun RGB-D Dataset

	Methods	Data	mAcc (%)	mIoU (%)
Methods	Lstm-cf[29]	RGB-D	48.1	-
	FuseNet[17]	RGB-D	48.3	37.3
	Qi et al.(VGG-16)[38]	RGB-D	55.2	42.0
	Wang et al.(VGG-16)[48]	RGB-D	53.5	42.0
	Depth-aware[47]	RGB-D	53.5	42.0
	Context[33]	RGB	53.4	42.3
	CFN(VGG-16)[30]	RGB-D	-	42.5
	RefineNet-Res101[31]	RGB	57.8	45.7
	RDF-152[28]	RGB-D	60.1	47.7
	RedNet[21]	RGB-D	60.3	47.8
	ACNet[19]	RGB-D	60.3	48.1
Ours	VCD+RedNet	RGB-D	**62.9**	**50.3**
	VCD+ACNet	RGB-D	**64.1**	**51.2**

Table 2. Performance Comparisons on NYUv2 Dataset

	Methods	Data	mAcc (%)	mIoU (%)
Methods	Eigen et al.[11]	RGB-D	45.1	34.1
	He et al.[18]	RGB-D	53.8	40.1
	Qi et al.(VGG-16)[38]	RGB-D	55.2	42.0
	D-CNN+HHA[47]	RGB-D	56.3	43.9
	Refine-101[31]	RGB	57.8	44.9
	Pixel Attention(ResNet-50)[26]	RGB-D	-	46.5
	CFN(RefineNet-152)	RGB-D	-	47.7
	Wang et al.(VGG-16)[48]	RGB-D	56.3	43.9
	RDFNet(ResNet-50)[28]	RGB-D	60.4	47.7
	RedNet[21]	RGB-D	62.6	47.2
	ACNet[19]	RGB-D	63.1	48.3
Ours	VCD+Deeplab(VGG16)	RGB-D	58.5	**45.3**
	VCD+RedNet(ResNet-50)	RGB-D	63.5	**50.7**
	VCD+ACNet	RGB-D	64.4	**51.9**

segmentation, mean pixel accuracy and mean intersection-over-union (mIoU) are adopted as the evaluation metrics.

4.2. Parameters Setup and Implementation Details

Since the proposed module can be plugged into existing deep networks readily, we choose to replace the convolution layers of existing networks with VCD module to show the effectiveness of our method. For RGB-D semantic segmentation, RedNet [21] and ACNet [19] are adopted as the baseline methods. As for the Cityscapes dataset, an excellent work HRNet [44] is employed as the baseline.

All the experiments are conducted on 4 NVIDIA GeForce Titan X Pascal GPU cards, and the batch size is set to 2 for each card. For RGB-D datasets, the input image size is set to 480×640. Stochastic Gradient Descent (SGD) optimizer is employed for all the experiments. The initial learning rate lr is set to 2e-3 for RGB-D dataset, while lr is set to 1e-2 for Cityscapes dataset. The learning rates for $conv_\mu$ and $conv_\sigma$ of VCD module are set to $10 \times lr$. For RGB-D datasets, 300 epochs are used for training the model, including 270 epochs for the first-stage training, and the last 30 epochs for the second stage. 484 epochs are used for the first training stage on Cityscapes dataset. 80 extra epochs are used for the second stage training. L in Eq. 6 is set to 5 in all the experiments.

4.3. Performance on SUN RGB-D Dataset

The comparison with several state-of-the-art methods is reported in Table 1. It can be seen that the performance of the proposed methods *VCD+RedNet* and *VCD+ACNet* outperform the compared state-of-the-art methods. Note that, no post-processing like CRF is used in all of our experiments. We simply replace the CNN layers with the VCD module in the last two stages of ResNet. By using the VCD module, *VCD+RedNet* can improve the mIoU from 47.8% to 50.3%. When choosing ACNet as our baseline, the performance can also be boosted significantly with the VCD module. As presented in Table 1, *VCD+ACNet* can achieve 51.2%, which largely improves the performance of the baseline method. These results demonstrate that learning adaptive receptive-field is useful for improving segmentation performance and the proposed method is effective for RGB-D scene semantic segmentation.

4.4. Performance on NYUv2 Dataset

Table 2 displays the performance comparison on NYU-v2 dataset. It can be seen that the proposed method obtains superior performance over other methods. The comparison results indicate that the proposed VCD module is an effective way for better exploiting the depth cues. By plugging the VCD module to the RedNet baseline, the proposed method can achieve a large performance improvement (3.5% mIoU). We attribute this gain to the enhanced deep features learned with adaptive-spatial context, which can maintain more image details and focus on small clutter objects.

The improvement on both ACNet and RedNet baseline indicates that the proposed method is orthonormal to the choice of the deep network architecture. By replacing standard CNN layer with VCD module, the performance can be boosted consistently. Depth-aware and Pixel-wise attention are the most related works to ours. To compare with Depth-aware CNN, i.e., *D-CNN+HHA*, we conduct experiments on Deeplab[2] baseline with VGG-16 backbone. The proposed *VCD+Deeplab* can achieve better performance than its counterpart *D-CNN+HHA*.

To get a more comprehensive understanding of the VCD module, the learned scale-guidance maps are visualize in Fig. 6. It can be seen that the learned receptive-fields are different for flat area like *wall* and objects like *toilet*, which is in accordance with our intuition. Moreover, the scale-guidance maps learned with multi-modality are also shown in Fig. 6. Compared with maps learned with only RGB modality, multi-modality guided scale-guidance maps are more reasonable.

4.5. Ablation Study

For a more comprehensive evaluation, we conduct experiments to study the effect of context-deformable convo-

Table 3. Ablation Study Results on NYUv2 Dataset

Method	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	blinds	desk	shelves	curtain	dresser	pillow	mirror	floormat	mIoU
DeeplabV3[1]	**78.8**	83.4	56.7	61.9	57.0	59.4	41.3	39.9	44.5	45.1	60.3	56.9	54.9	22.9	14.2	52.4	40.6	40.1	31.3	30.8	43.8
DCN[8]	77.0	83.0	56.4	64.7	57.0	60.8	39.9	35.5	44.6	44.7	59.3	55.8	**59.9**	20.3	12.3	55.9	**51.2**	39.8	36.2	34.2	44.9
DCNV2[56]	76.8	82.7	56.3	64.3	57.0	59.9	39.9	36.3	44.5	45.6	60.3	56.0	57.1	21.1	**16.6**	54.1	47.3	41.9	36.5	32.8	45.1
VCD	78.2	83.7	57.4	66.1	57.2	60.9	40.1	39.5	45.1	**46.8**	59.4	58.1	56.6	21.9	16.0	55.2	47.0	42.7	36.2	34.3	46.0
VCD+DCN	77.9	83.2	**59.3**	66.3	58.5	59.4	43.4	**42.1**	**48.8**	44.5	59.9	58.3	58.1	**23.9**	16.7	53.9	47.7	44.9	38.3	30.6	46.4
VCD††	78.6	**86.3**	58.1	**73.5**	**60.5**	**63.7**	**44.1**	36.0	44.2	44.3	**60.7**	**62.5**	58.7	20.4	15.9	**63.6**	50.8	**48.9**	**41.8**	**35.0**	**47.1**

Method	clothes	ceiling	books	fridge	tv	paper	towel	shower	box	board	person	nightstand	toilet	sink	lamp	bathtub	bag	ot.struct.	ot.furn.	ot.props.	mIoU
DeeplabV3	20.7	69.8	30.3	42.8	52.5	27.7	33.2	24.5	13.6	68.9	73.3	37.7	65.1	51.3	39.2	36.4	**12.5**	27.7	15.2	36.6	43.8
DCN[8]	22.3	63.3	26.9	52.8	58.7	29.9	**39.8**	40.4	14.9	65.3	76.2	39.9	67.1	50.3	38.7	40.1	7.3	26.7	16.5	36.9	44.9
DCNV2[56]	21.0	64.2	29.1	**54.3**	**60.2**	27.8	38.1	39.7	13.0	65.5	76.1	43.6	66.2	48.8	40.0	40.6	10.1	26.9	16.2	36.0	45.1
VCD	22.2	67.0	30.0	50.9	57.0	30.7	36.7	40.6	**15.6**	**72.6**	**77.5**	41.2	69.1	51.8	43.0	39.4	9.5	27.7	18.3	37.0	46.0
VCD+DCN	20.8	64.6	**33.6**	51.2	57.9	**33.9**	37.5	40.0	15.2	66.8	76.3	38.8	74.2	52.9	**45.4**	36.5	11.5	**28.8**	15.5	**37.2**	46.4
VCD††	**24.1**	**70.4**	26.9	48.6	57.9	23.9	38.1	**45.6**	8.3	44.0	70.6	**48.2**	**75.6**	**60.2**	43.7	**55.0**	11.6	26.0	**18.4**	36.6	**47.1**

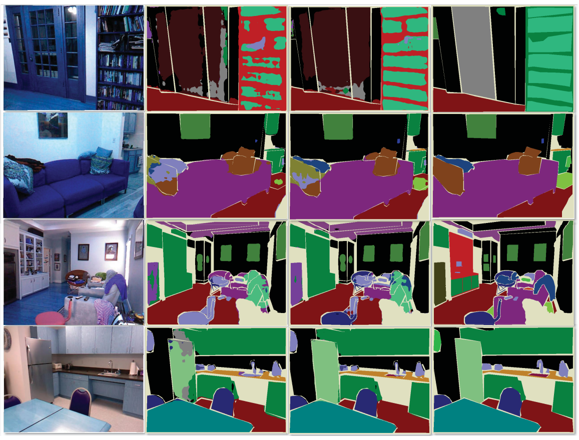

Figure 5. Visualization of segmentation results on NYUv2 dataset. For each row, the images are 1) the input; 2) the result of DCNv2; 3) the result of the proposed method; 4) the ground truth.

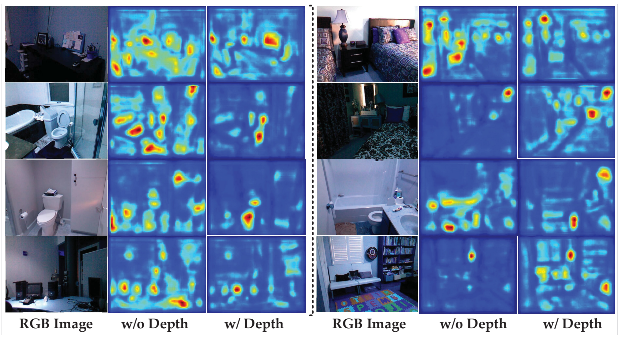

Figure 6. Visualization of the last scale-guidance map on NYUv2 dataset. We show the map with RGB modality and with multi-modal modality. The left image triplets are results using only the RGB modality. The right ones are the scale-guidance map using multi-modality information. (Best viewed in color.)

lution and variational Bayesian modeling. The results are presented in Fig. 3. Additionally, as VCD module is integrated with Deformable ConvNets, we also compare the proposed method with DCN and DCNv2. From the results we can see that the baseline (DeeplabV3) method achieves 43.8% mIoU. DCN can improve the baseline from 43.8% to 44.9%, and DCNv2 can also boost the performance from 43.8% to 45.1%. We attribute this performance gain to the deformable spatial context. VCD denotes the VCD module that is not integrated with DCN. When not integrated with DCN, VCD can outperform the baseline methods clearly. Nevertheless, when VCD is integrated with DCN, VCD+DCN can take advantage of both the scale and shape deformation to further enhance the segmentation performance. VCD†† denotes the VCD module trained with DCN and the stochastic mechanism. VCD†† can further boost the performance to 47.1% mIoU compared with VCD†. This indicates that it is useful to model the scale-guidance map in a Bayesian probabilistic framework. Note that only RGB images are used in the experiments presented in Table 3. Some qualitative results are shown in Fig. 5.

4.6. Performance on Cityscapes Dataset

Since the proposed VCD module can also be applied to the RGB image based semantic segmentation task, we also study the effect of it on street scene RGB segmentation task. In this part, experiments are conducted on a widely-used segmentation benchmark: Cityscapes. We choose HRNet [44] as our baseline, and replace the last four branches of stage 4 with VCD module. Our method has been evaluated on the cityscapes benchmark, the detailed results can be seen at this link.

https://www.cityscapes-dataset.com/method-details/?submissionID=5713

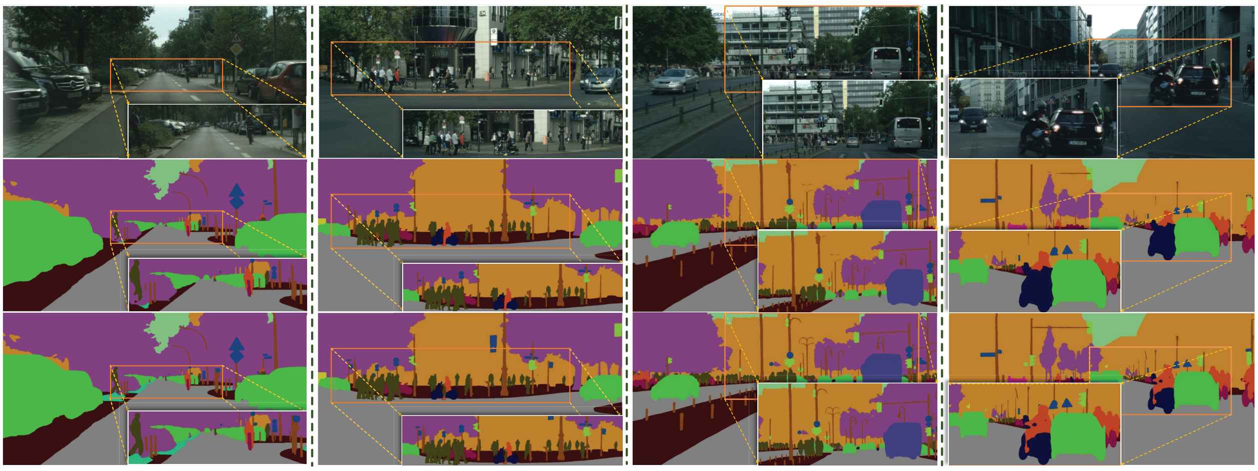

Figure 7. Comparison results on Cityscapes *test* set. Input images are in the first row. The qualitative results of baseline and the proposed method are displayed in the 2^{nd} and the 3^{rd} row respectively.

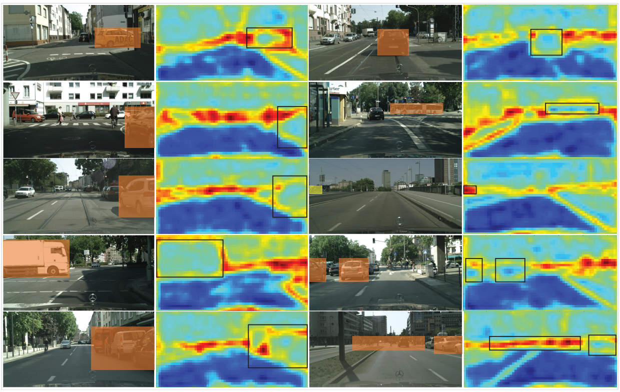

Figure 8. Visualization of the last scale-guidance map on Cityscapes. It is obvious that the learned scale-guidence map is reasonable.

As shown in Table 4, by simply replacing standard convolution with the proposed VCD module, our method can achieve 82.3% mIoU on Cityscapes, which outperforms the compared state-of-the-art methods. More specifically, Our method ranks 3^{rd} on IoU class metric and ranks 2^{nd} on iIoU category metric when coarse data or other dataset is not used. The scale-guidance map is presented in Fig. 8. Obviously, it can be seen that the scale-guidance maps are essentially different from attention maps. As shown in Fig. 8, the receptive-field for cars with large scale is quite different from small cars. Thus, the learned map is not focusing on objects, which has essential difference with self-attention maps. It has been trained to predict spatial context correlated maps for controlling the receptive-field adaptively.

Table 4. Performance Comparisons on Sun Cityscapes **test** Dataset

Method	Validation	Coarse	Backbone	mIoU(%)
PSPNet[54]	×	×	ResNet-101	78.4
PSANet[55]	×	×	ResNet-101	78.6
Scale-adaptive[52]	✓	×	ResNet-101	78.1
AAF[23]	x	×	ResNet-101	79.1
BiSeNet[23]	✓	×	ResNet-101	78.9
PSANet[55]	✓	×	ResNet-101	80.1
DFN[50]	✓	×	ResNet-101	79.3
DANet[12]	✓	×	ResNet-101	81.5
PSPNet[54]	✓	✓	ResNet-101	81.2
DenseASPP[49]	✓	×	DenseNet-161	80.6
DeepLabv3[1]	✓	✓	ResNet-101	81.3
HRNet[44]	✓	×	HRNetV2-W48	81.6
OCR[51]	✓	×	ResNet-101	81.8
Ours(VCD)	✓	×	HRNetV2-W48	**82.3**

5. Conclusion

In this paper, a variational context-deformable (VCD) module is proposed to learn adaptive spatial-context in a structured manner. The VCD module learns a deformable spatial-context with the guidance of RGB and Depth modality information. Specifically, adaptive Gaussian kernels are learned with the variational technique and the guidance of multi-modal information. By multiplying the learned Gaussian kernel with fixed receptive-size, the VCD module can aggregate flexible spatial context for each pixel during convolution. Experiments have demonstrated the effectiveness of the proposed method.

6. Acknowledgment

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61632018, 61825603, U1864204 and 61773316.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[3] Yunlu Chen, Thomas Mensink, and Efstratios Gavves. 3d neighborhood convolution: Learning depth-aware features for RGB-D and RGB semantic segmentation. In *2019 International Conference on 3D Vision, 3DV 2019, Québec City, QC, Canada, September 16-19, 2019*, pages 173–182. IEEE, 2019.

[4] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition7*, pages 1475–1483, 2017.

[5] Hang Chu, Wei-Chiu Ma, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Surfconv: Bridging 3d and 2d convolution for RGBD images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3002–3011. IEEE Computer Society, 2018.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Camille Couprie, Cl{é}ment Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 764–773, 2017.

[9] Zhuo Deng, Sinisa Todorovic, and Longin Jan Latecki. Semantic segmentation of RGBD images with mutex constraints. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 1733–1741, 2015.

[10] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019.

[11] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2650–2658, 2015.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3146–3154, 2019.

[13] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149, 2015.

[14] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[15] Saurabh Gupta, Pablo Arbelez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[16] Saurabh Gupta, Ross Girshick, Pablo Arbelez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. 8695:345–360, 2014.

[17] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. pages 213–228, 2016.

[18] Yang He, Wei-Chen Chiu, Margret Keuper, and Mario Fritz. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for RGBD semantic segmentation. *CoRR*, abs/1905.10089, 2019.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *CoRR*, abs/1811.11721, 2018.

[21] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor RGB-D semantic segmentation. *CoRR*, abs/1806.01054, 2018.

[22] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson W. H. Lau, and Thomas S. Huang. Geometry-aware distillation for indoor semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2869–2878, 2019.

[23] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X. Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, pages 605–621, 2018.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[25] Shu Kong and Charless C. Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Shu Kong and Charless C. Fowlkes. Pixel-wise attentional gating for scene parsing. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1024–1033, 2019.

[27] Hema Swetha Koppula, Abhishek Anand, Thorsten Joachims, and Ashutosh Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *International Conference on Neural Information Processing Systems*, pages 244–252, 2011.

[28] Seungyong Lee, Seong Jin Park, and Ki Sang Hong. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[29] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. pages 541–557, 2016.

[30] Di Lin, Guangyong Chen, Daniel Cohenor, Pheng Ann Heng, and Hui Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *IEEE International Conference on Computer Vision*, pages 1320–1328, 2017.

[31] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

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

[33] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[34] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4898–4906, 2016.

[35] Carlo Dal Mutto, Pietro Zanuttigh, and Guido M. Cortelazzo. Fusion of geometry and color information for scene segmentation. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):505–521, 2012.

[36] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012.

[37] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[38] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *CVPR*, pages 5199–5208, 2017.

[39] E Shelhamer, J. Long, and T Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2014.

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. 7576(1):746–760, 2012.

[41] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[42] Luciano Spinello and O. Arras Kai. People detection in rgb-d data. In *Ieee/rsj International Conference on Intelligent Robots and Systems*, pages 3838–3843, 2011.

[43] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik G. Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11166–11175, 2019.

[44] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[45] Camillo J Taylor and Anthony Cowley. Segmentation and analysis of rgb-d data. In *RSS 2011 workshop on RGB-D cameras*, volume 90, 2011.

[46] Karthik Mahesh Varadarajan and Markus Vincze. Object part segmentation and classification in range images for grasping. In *Advanced Robotics (ICAR), 2011 15th International Conference on*, pages 21–27. IEEE, 2011.

[47] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, pages 144–161, 2018.

[48] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. *arXiv preprint arXiv:1803.06791*, 2018.

[49] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3684–3692, 2018.

[50] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1857–1866, 2018.

[51] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *CoRR*, abs/1909.11065, 2019.

[52] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2050–2058, 2017.

[53] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012.

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[55] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Computer Vision - ECCV 2018 - 15th European Conference, Mu-*

nich, Germany, September 8-14, 2018, Proceedings, Part IX, pages 270–286, 2018.

[56] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.

[57] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. *CoRR*, abs/1908.07678, 2019.