Gated Path Selection Network for Semantic Segmentation

Qichuan Geng, Hong Zhang, Xiaojuan Qi, Gao Huang, *Member, IEEE,*
Ruigang Yang, *Senior Member, IEEE,* Zhong Zhou, *Member, IEEE,*

Abstract—Semantic segmentation is a challenging task that needs to handle large scale variations, deformations, and different viewpoints. In this paper, we develop a novel network named Gated Path Selection Network (GPSNet), which aims to adaptively select receptive fields while maintaining the dense sampling capability. In GPSNet, we first design a two-dimensional SuperNet, which densely incorporates features from growing receptive fields. And then, a Comparative Feature Aggregation (CFA) module is introduced to dynamically aggregate discriminative semantic context. In contrast to previous works that focus on optimizing sparse sampling locations on regular grids, GPSNet can adaptively harvest free form dense semantic context information. The derived adaptive receptive fields and dense sampling locations are data-dependent and flexible which can model various contexts of objects. On two representative semantic segmentation datasets, *i.e.*, Cityscapes and ADE20K, we show that the proposed approach consistently outperforms previous methods without bells and whistles.

Index Terms—Semantic Segmentation, Local Discriminative Feature, Adaptive Context Aggregation, Adaptive Receptive Fields and Sampling Locations.

1 Introduction

Semantic segmentation refers to the problem of assigning a semantic object category to each pixel. Recent progress in semantic segmentation [1], [2], [3], [4], [5], [6], [7] largely benefits from Deep Convolutional Neural Networks (DCNNs) [8], [9], [10]. However, DCNNs are inherently limited by the manually defined structures, where the receptive fields are restricted to constant regions [1], [2], [3], [11]. In contrast, objects in images are in a large range of scales, deformations and viewpoints, and thus the unchangeable receptive fields in CNNs are insufficient to deal with appearance variations.

Extensive efforts have been made to enlarge and enrich receptive fields to better understand the semantic scenes [1], [2], [3], [4], [5], [11]. Atrous convolution [2] incorporates larger contexts by dilating the convolution kernel in a fixed manner, which lacks the ability to cope with multi-scale objects. To mitigate the problem, PSPNet [6] applies *pyramid pooling module* to aggregate information from different scales of feature maps. ASPP [3] and DenseASPP [7] are introduced to use a series of atrous convolution layers to learn features with multiple dilation rates. Moreover, in order to recover the spatial information, features of multiple semantic levels from backbone networks are combined to obtain appropriate representations [1], [5], [12], [13]. Nevertheless, the above approaches suffer from the common issues – the receptive fields are all regular and determinate, which may be incapable to handle objects with various forms and sizes. Meanwhile, unsuitable receptive fields may degenerate the representative capability or ignore important details due to sparse sampling with dilation [4], [14], [15].

Further, to capture rich semantic context, attention-based approaches [16], [17], [18], [19] are proposed to adaptively aggregate short- and long-range features. As incorporating more sampling locations, the global context aggregation helps eliminate the confusion of pixel-wise classification. To highlight local discriminative information, recent works [20], [21], [22] show that the adaptive sampling locations can be acquired by predicting additional offsets. However, the learned receptive fields can only sparsely sample a fixed number of locations rather than considering the overall relevant contexts, which can leverage rich spatial details for semantic segmentation [7], [15].

Although adopting the adaptive receptive fields or dense contexts helps improve semantic segmentation by a large margin, the aforementioned methods are still not comprehensive solutions. To be specific, these models rely on either designing special network architectures or developing different sampling strategies like increasing the sampling locations or the number of samples. In this paper, we propose *Gated Path Selection Network* (*GPSNet*) to learn adaptive receptive fields and select dense samples for semantic segmentation. As shown in Fig. 1, GPSNet is able to harvest various contextual information via densely aggregating features from small to large receptive fields. The adaptive aggregation method is data-dependent and flexible to model various scales, geometric deformations and different viewpoints.

Specifically, the GPS module consists of two components, *i.e.*,

- *Qichuan Geng and Zhong Zhou are with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China. This work was partially done when Qichuan Geng was an intern at Baidu Research. E-mail: zhaokefirst@buaa.edu.cn, zz@buaa.edu.cn.*
- *Hong Zhang is with National Engineering Laboratory of Deep Learning Technology and Application, China, and also with Baidu Research. E-mail: fykalviny@gmail.com.*
- *Ruigang Yang is with University of Kentucky and also with Inceptio Technology. This work was done in Baidu Research. E-mail: ryang2@uky.edu.*
- *Xiaojuan Qi is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. E-mail: qxj0125@gmail.com.*
- *Gao Huang is with the Department of Automation, Tsinghua University, Beijing, China. E-mail: gaohuang@tsinghua.edu.cn.*
- *Corresponding author: Zhong Zhou.*

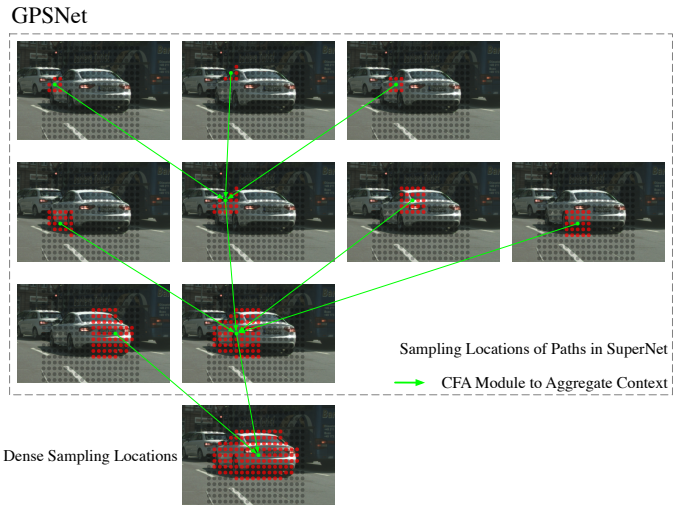


Fig. 1. Aiming at generating desirable representation for different scales and geometric variations, *GPSNet* is proposed to densely select context for objects to promote recognition. It can be observed that the target sampling locations can be recursively decomposed into the union of sampling location sets. To this end, *SuperNet* is designed as a two-dimensional network to provide adequate paths, and *CFAs* are further introduced to gradually select paths, where the receptive fields arrange from small to large.

SuperNet, and Comparative Feature Aggregation (CFA) module. SuperNet is a two-dimensional network equipped with horizontal and vertical connections, which accommodates various sub-pyramids, to maintain alterable multi-scale features. To facilitate the learning of adaptive local discriminative representations, another core component CFA is further introduced to aggregate features within SuperNet. CFA predicts soft masks to regularize receptive fields, which allows dynamical selections of effective locations, promoting desirable representation generation for different scales and geometric variations.

Finally, we summarize our main contributions of this paper as follows:

1) We are the first to explicitly consider *receptive fields*, *sampling locations* and *the number of sampling locations* simultaneously in semantic segmentation.

2) *GPS module* is proposed to efficiently aggregate discriminative contexts. Such a module is model-agnostic that can be readily used in various ASPP-like structures and trained in an end-to-end manner.

3) Extensive experimental results on Cityscapes [23], and ADE20K [24] demonstrate that *GPSNet* consistently improves the performance of previous state-of-the-art approaches.

More details are given in the following sections. In Section 2, we review previous works. We describe the GPSNet architecture and its analysis in Section 3 and 4 respectively. In Section 5, we evaluate the performance of GPSNet on two semantic segmentation datasets. This is followed by a discussion regarding our approach. We conclude in Section 6.

2 Related Work

Pixel-wise semantic segmentation task has been largely driven by deep fully convolutional neural networks (FCNs) [1], [3], [6], [25], [26]. Since the pioneering work FCN [1], recent works have shown that contextual information is important for improving semantic segmentation accuracy. The contextual information in segmentation help not only to reduce the ambiguity in recognition but also to avoid overwhelming by other salient objects [6], [14], [17], [18].

Receptive fields in semantic segmentation. Deeplab [2] and Dilated Conv [26] proposed the atrous convolution to enlarge the network receptive field without sacrificing the resolution which recently became popular in semantic segmentation. Atrous convolution enables the network to harvest contextual information in a larger region for semantic segmentation. To capture long-range contexts in the downsampled feature maps, symmetric, separable large filters [11] can be adopted to reduce the model parameters and computation cost. The aggregations of contextual information from hand-engineered fixed regions are still limited for modeling large contextual variations in objects. Yunho *et al*. [22] proposed active convolution with learnable offsets to provide greater freedom to form CNN structures. The deformable convolution layer was introduced in [20], [21], which makes the convolution kernel adaptive to geometric variations of the object, extracting dynamic contextual information for image recognition.

Spatial details in semantic segmentation. In contrast to the challenge in image classification, which recognizes the dominant objects in the whole image, semantic segmentation needs to assign object category to each pixel, especially to localize the details. Different feature aggregation strategies help improve the discrimination of local features [3], [7], [20], [21].

Local features can be obtained by processing different scales of feature maps. Centered on the target pixel, those local features corresponding to different receptive fields and semantic levels are aggregated to recover spatial details. To facilitate dense prediction, the deconvolutional layer was also introduced in [1], [27], which is a learnable upsampling operation. With multi-level feature maps, FCN [1] added shallow predictions into upsampled predictions multiple times to recover the resolutions and the details. The following work SegNet [28] introduced an encoder and decoder network, where the decoder utilizes pooling indices in the encoding layers to upsample the feature map. Further, to develop real-time semantic segmentation networks for practical applications, Paszke *et al*. proposed the light-weight ENet [29] by exploiting a small decoder to fine-tune the details. To refine segment contours, CRF was applied as a post-processing procedure [3] or end-to-end integrated [25] into the network. Zhao *et al*. [12] proposed ICNet to utilize the image pyramid to optimize the network, which adaptively derived semantic segmentation results from a lower resolution to higher resolution stage by stage to balance the performance and efficiency of semantic segmentation.

Multi-scale feature extraction in semantic segmentation. As most state-of-the-art feature encoders are pre-trained on ImageNet [30] for image recognition, the feature corresponding to a fixed receptive field could be infeasible for different forms of objects in semantic segmentation.

Multi-scale features corresponding to different nested regions have been proved effective to improve semantic segmentation and robustness to scale variations. UNet [13] adopted skip connections to combine shallow representations from the encoder and deep features from the decoder, which exploit low-level features for accurate semantic segmentation. Moreover, an atrous spatial pyramid pooling (ASPP) [3], [4], [5] module was widely employed to incorporate contextual information from multiple scales. Meanwhile, the low-level features also can be incorporated to refine

the segmentation results along object boundaries [5]. [31] made use of feature maps at different scales to produce weight maps, which are applied to weighted merge score maps at different scales respectively. MSCI [32] adopted LSTM to intertwine bidirectional contexts of super-pixels in multi-scale feature maps. Ding *et al*. [14] proposed a context contrasted local model to construct multi-scale and multi-level context contrasted local features.

Dense sampling in semantic segmentation. DenseASPP [7] organized atrous convolutional layers in a cascade fashion to provide a denser feature pyramid and a larger receptive field, which involved more locations in the computation. To eliminate the gridding effect introduced by atrous convolutions, ResNet-DUC-HDC model [15] assigned adjusted dilation rates to a serial of convolutional layers.

In order to capture the global context, scene recognition methods [6], [17], [33] and comparisons between semantic objects/stuffs [16], [18], [19], [34] are introduced. ParseNet [33] proposed the global average pooling layer which introduced global contextual information for semantic segmentation. Later, Zhao *et al*. [6] proposed a Pyramid Pooling Module to aggregate contextual information from multi-scale regions. More recently, Point-wise Spatial Attention Network [16] selected information through a learned attention map to dynamically adjust contextual information aggregation.

Attention-based mechanisms enable the network to adaptively select the context for each location. Chen *et al*. [31] proposed to learn combination weights for composing multi-scale features. OCNet [18] introduced an object context network to learn an object context map by modeling pixel-pixel similarities which are further utilized to refine the representations of each pixel. CCNet [34] harvested the contextual information on the criss-cross path which can provide long-range contextual information to each pixel with improved efficiency. Wang *et al*. [35] proposed a non-local operation that computes a weighted sum of features in the global map based on the attention mechanism.

Li *et al*. [36] introduced expectation-maximization to estimate a compact set of bases and low-rank attention maps. It demonstrates that point-wise comparison between features and representative information derived from the feature map help produce attention maps. More recently, attention-based context is proved to serve as a global constraint as the query positions modeled by non-local networks almost share the same contexts [37], [38].

Path selection in semantic segmentation. Auto-Deeplab [39] extended the idea of neural architecture search to optimize a network-level structure. Dynamic routing [40] generated forward paths for each location on-the-fly without searching, which can be trained in an end-to-end manner. However, these architecture search methods are hard to benefit from well-trained backbones. With Gated Fully Fusion (GFF) module, GFFNet [41] enhanced features of different blocks, where high-level features are with strong semantic context and low-level features are with more details. GSCNN [42] proposed the gated convolutional layer (GCL) to gate the lower-level activations in the shape stream with the higher-level activations in the semantic stream.

In parallel to recent works that emphasize more on capturing global context, our approach focuses on extracting rich local features. Our proposed adaptive GPSNet allows densely selecting samples to generate more discriminative local features. Experimental results demonstrate that our method can work together with global-based methods to further boost performance.

3 Gated Path Selection Network

In this section, we first describe the overall framework of *Gated Path Selection Network* (*GPSNet*)), then we introduce how to dynamically learn discriminative representation with *Gated Prediction Selection* (*GPS*) module.

3.1 GPSNet Framework

The overall framework of GPSNet is shown in Fig. 2. It builds on a fully convolutional architecture (ResNet-101) pre-trained on ImageNet. We use dilated convolution layers to maintain the resolutions of feature maps. The GPS module in Fig. 2 extracts discriminative local information built on a SuperNet to generate a set of multi-scale features. Furthermore, the Comparative Feature Aggregation (CFA) is introduced to aggregate the features in a flexible way.

We elaborate the GPS module and the corresponding design in the following.

3.2 GPS Module

Understanding and utilizing the contextual information is of vital importance in semantic segmentation. The backbone network pre-trained on ImageNet has provided high-level semantic features. However, the features are far from optimal due to the following two aspects: 1) the learned features are dominated by the salient objects which are incapable to recognize different objects or stuff; 2) the features share the same receptive fields, making it difficult or infeasible for complex geometric transformations.

To address the issues, many existing works devote to obtain informative local context [3], [14]. Differently, we propose a GPS module to select relevant context jointly by considering *receptive fields*, *sampling locations* and *the number of sampling locations*. As shown in Fig. 2, GPS module is built on SuperNet, which keeps constant feature channels to reduce the computational cost and balances the channel proportion for different scales. Further with CFA, samples from larger receptive fields participate in extracting discriminative local contexts. Instead of the vanilla concatenation in ASPP, we gather features with *Comparative Feature Concatenation*, which is a variant of CFA. Via point-wise aggregation, GPS module can gradually enhance the features.

3.2.1 SuperNet

To enhance the capability of learning effective feature representations, we propose a two-dimensional network–*SuperNet*, to improve ASPP-like network structure with multiple entrances and exits by propagating information among branches. To ensure both the interior and exterior regions can be densely sampled, we improve the original ASPP structure with three techniques: *Tuned Dilation*, *Bottlenecked Branch*, and *Dense Connectivity*.

ASPP-like Structure. Atrous Spatial Pyramid Pooling (ASPP) is proposed to concatenate feature maps from multiple parallel atrous convolution layers with different dilation rates. For each branch, fixed locations at different intervals are sparsely sampled. However, as the dilation rate increasing, the atrous convolution layer is to lose the capability to capture information effectively [4]. Meanwhile, the atrous convolution layers with large receptive fields tend to ignore several details due to the large interval and sparse sampling.

Tuned Dilation. We first extend the parallel atrous convolution layers in ASPP to a grid form. The atrous convolution layers with

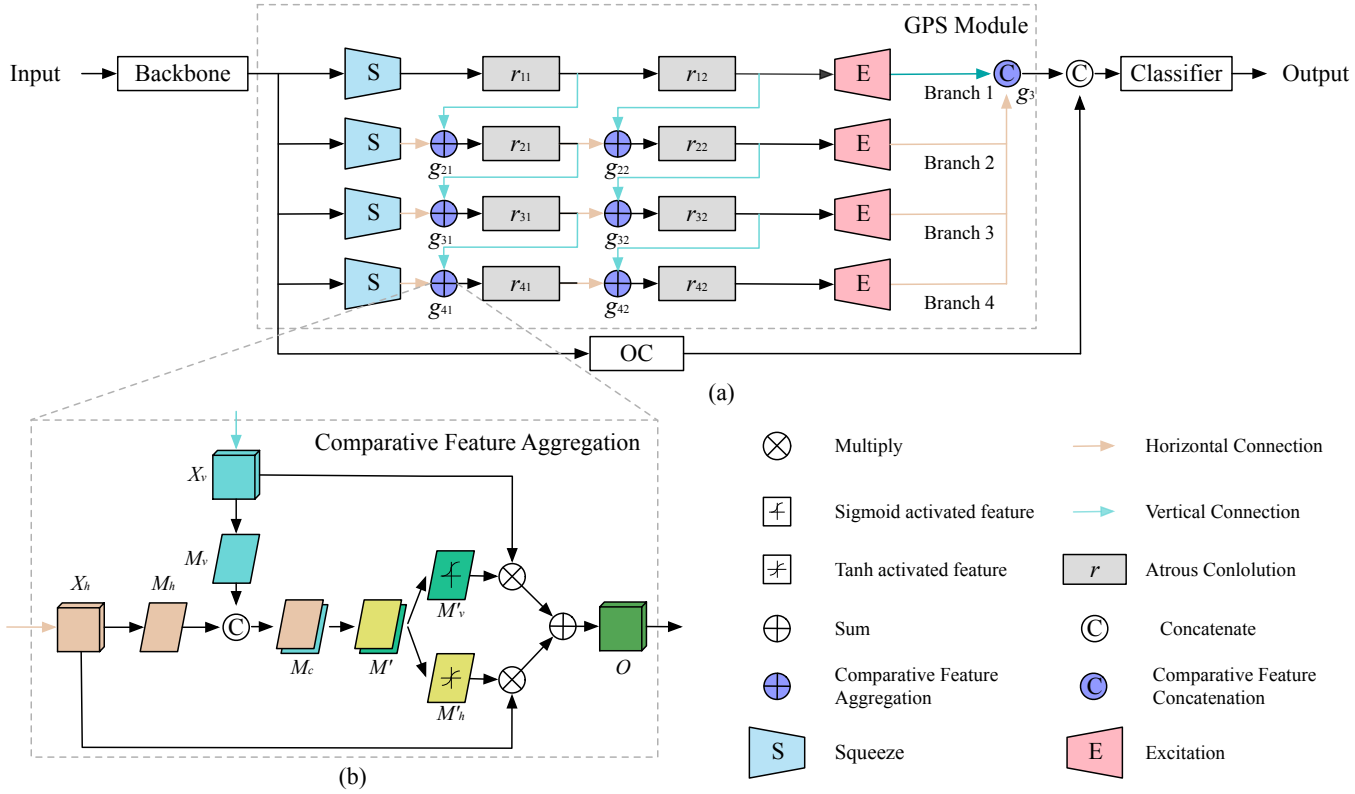


Fig. 2. (a) **Overview of GPSNet**. Given an input image, a 1×1 convolution layer is used to squeeze the feature maps. Then we use atrous convolution layers with different dilation rates r_{ij} to learn adaptive receptive fields to sample features. At last, the features are excited by a 1×1 convolution layer. To get global context-aware features, we further integrate the Object Context (OC) module into our network. (b) **An illustration of *CFA***. It takes the vertical features X_v and horizontal features X_h as inputs, and estimates the soft masks $\{M'_v, M'_h\}$to reweight the original features by element-wise multiplication. Final features are produced by summing the reweighted features. Consistent with ASPP, we aggregate different branches with *Comparative Feature Concatenation*, which are constructed by replacing sum with concatenation in the *CFA*.

dilation rates $\{r_1, r_2, r_3, r_4\}$ are doubled into untuned grid-form dilation rates $\{(r_1, r_1), (r_2, r_2), (r_3, r_3), (r_4, r_4)\}$. To mitigate repeatedly sampling, the dilation rates are further tuned to improve the sampling rate. As illustrated in Fig. 2(a), we replace the dilation parameters with prime numbers which are $\{(1, 3), (11, 13), (23, 29), (33, 37)\}$ to produce tuned dilation rates $\{(r_{11}, r_{12}), (r_{21}, r_{22}), (r_{31}, r_{32}), (r_{41}, r_{42})\}$.

Bottlenecked Branch. To alleviate the computing resource overhead especially for the GPU memory usage, inspired by eASPP [43], we introduce bottlenecked branches in SuperNet. Following [44], each of the branches starts with a *Squeeze* operation to reduce the channel of input features with a 1×1 convolution. Then two consecutive 3×3 atrous convolutions extract features with different sampling rates. Finally, an *Excitation* operation is applied at the exit of each branch, the features are expanded to large channel features with a 1×1 convolution. All the convolutions are followed by InplaceABNsync [45].

Dense Connectivity. To facilitate information flow across atrous convolution layers, we use dense connectivity [46] to bridge parallel bottlenecked branches in SuperNet. The intermediate feature maps are aggregated by the results of one or two directions: 1) the output of a convolution from the previous layer in the same branch (horizontal connection) and, if possible, 2) the result of a convolution from the previous branch (vertical connection). Consequently, the subsequent layers gather the information from early layers with relatively small receptive fields. In comparison with ASPP and DenseASPP, because of the grid-form dense connectivity pattern, we can acquire abundant features with more diverse and denser contexts.

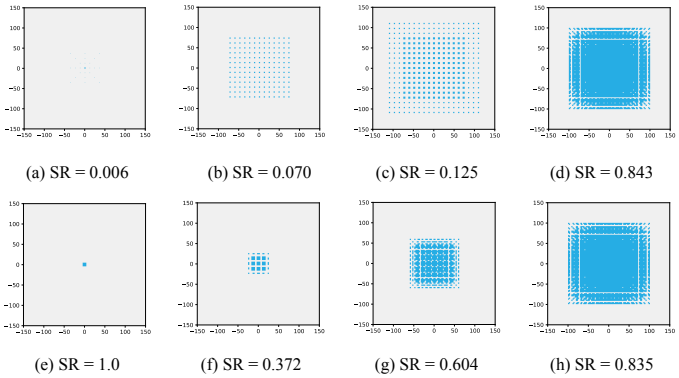
3.2.2 Comparative Feature Aggregation

In this subsection, we aim to dynamically aggregate the rich scale features in SuperNet. In existing works, different features are usually merged through direct concatenation and summation. In fact, objects are in complex geometric transformations, augmented features via simple concatenation or summation may be infeasible in dense prediction task. In CFA, features from the vertical and horizontal directions are compared and fused to produce discriminative contexts. Fig. 2 (b) depicts the process of CFA. It contains three operators: *Projection*, *Comparison* and *Weighted Sum/Concatenation*.

Projection. As the changes between contexts in neighboring receptive fields are smooth, *Projection* produces the highly compressed indicator. It considers two input feature maps X_v, X_h with the shape of $H \times W \times C$, where X_v and X_h are the previous vertical feature maps and horizontal feature maps, respectively. The soft gate masks M_v, M_h with the size of $H \times W \times 1$ are predicted via a projection transformation $\mathcal{F}$, where

$$\mathcal{F} : X_i \rightarrow M_i, i \in \{v, h\}. \tag{1}$$

The transformation layer is defined with three consecutive operations: a 1×1 convolution, followed by a batch normalization (BN) and a rectified linear unit (ReLU).



(a) SR = 0.006 (b) SR = 0.070 (c) SR = 0.125 (d) SR = 0.843

(e) SR = 1.0 (f) SR = 0.372 (g) SR = 0.604 (h) SR = 0.835

Fig. 3. Illustration of sampling locations of (a) ASPP, (b) DenseASPP, (c) Untuned *GPSNet*, (d) *GPSNet*. It can be clearly observed that *GPSNet* more densely samples the locations in the receptive field. We further visualize the sampling locations of different branches of the *GPSNet*. In (e) Branch 1 of (d), (f) Branch 2 of (d), (g) Branch 3 of (d) and (h) Branch 4 of (d), the receptive fields of all the branches are regularized, and locations are also densely sampled.

Comparison. A *Comparison* is introduced to compare different projected features to encourage information incorporation. It predicts soft masks to point-wise integrate information from feature maps. Specifically, we first get the concatenated mask $M_c = [M_v, M_h]$, then a comparison function $\mathcal{C}_a$ is applied to get the soft mask $M' \in \mathbb{R}^{H\times W\times 2}$:

$$\mathcal{C}_a : M_c \rightarrow M'. \tag{2}$$

The comparison function $\mathcal{C}_a$ in our experiments consists of a 1×1 convolution layer, followed by a BN layer.

Weighted Sum/Concatenation. To aggregate the informative features, we first split M' into $\{M'_v, M'_h\}$ along the channel dimension, and re-weight the input feature $\{X_v, X_h\}$ by element-wise multiplication. The adaptive features $O \in \mathbb{R}^{H\times W\times C}$ are obtained by summing or concatenating the re-weighted features.

$$\begin{aligned} O_{\text{sum}} &= \text{sum}(\text{sigmoid}(M'_v) \otimes X_v, \tanh(M'_h) \otimes X_h), \\ O_{\text{concat}} &= \text{concat}(\text{sigmoid}(M'_v) \otimes X_v, \tanh(M'_h) \otimes X_h). \end{aligned} \tag{3}$$

Particularly, M'_h is used to select informative or redundant features from relatively large receptive fields, so we activate it with tanh. The motivations for this manipulation are: 1) add: accumulating the context of large receptive fields to collect long range information; 2) minus: removing the redundant features and spotlighting the local information in contrast to the context. Also, we want to preserve the previous information from small receptive fields, so we activate M'_v with sigmoid.

4 Model Analysis

In this section, we first introduce how to quantify the size of receptive fields and the ratio of maximum possible sampling locations (sampling rate) of several correlated models. Besides, we compare *GPSNet* with other existing works to further demonstrate our advantages.

TABLE 1
Comparing *GPS* module with other methods with the comparable dilation rates setting. *GPS* module provides larger receptive fields (RF), higher sampling rates (SR). To validate the effectiveness of the tuned dilation rate, we compare the results of the *GPS* module using tuned dilation rates (*GPSNet*) to skipping the tuned dilation technique (*Untuned GPSNet*).

Method	Dilation Setting	RF	SR
ASPP	{1,12,24,36}	73	0.006
DenseASPP	{1,12,24,36}	147	0.070
SuperNet	{(1, 1),(12, 12),(24, 24), (36, 36)}	219	0.125
Untuned *GPSNet*	{(1, 1),(12, 12),(24, 24), (36, 36)}	219	0.125
GPSNet	{(1, 3),(11, 13),(23, 29), (33, 37)}	199	0.843

4.1 Preliminary

In the following, we detail the size of receptive fields and sampling rates of several popular methods.

Atrous Convolution. The receptive field and sampling rate of atrous convolution are defined in Eqn. 4, where r and k are the dilation rate and kernel size respectively.

$$\begin{aligned} \text{RF}_{\text{ac}} &= (r \times k - r + 1)^2, \\ \text{SR}_{\text{ac}} &= \frac{k^2}{\text{RF}_{\text{ac}}}. \end{aligned} \tag{4}$$

ASPP. The receptive fields and the maximum possible number of sampling locations of ASPP can be obtained by overlaying atrous convolutions in different branches. Specifically, the receptive field and sampling rate are defined in Eqn. 5, where B is the number of branches, and b indicates the index of the branch.

$$\begin{aligned} \text{RF}_{\text{aspp}} &= (max(r_b) \times (k-1) + 1)^2, \\ \text{SR}_{\text{aspp}} &= \frac{(B \times k^2 - B + 1)}{\text{RF}_{\text{aspp}}}. \end{aligned} \tag{5}$$

Deformable Convolution Network (DCN) [20]. Different from convolution operators with fixed receptive fields, DCN adaptively samples k^2 locations, where k is the pre-defined deformable kernel size. Hence, RF_{dcn} can be approximated by the tightest bounding box surrounding all the sampling locations. Eqn. 6 gives the definition of receptive field and sampling rate of DCN:

$$\begin{aligned} \text{RF}_{\text{dcn}} &= (max(p_{i,x}) - min(p_{i,x})) \\ &\quad \times (max(p_{i,y}) - min(p_{i,y})), \\ \text{SR}_{\text{dcn}} &= \frac{k^2}{\text{RF}_{\text{dcn}}}, \end{aligned} \tag{6}$$

where $(p_{i,x}, p_{i,y})$ is the position of the i^{th} sample.

Context Contrasted Local Feature (CCL) [14]. CCL spotlights the local information by removing context from the relative large receptive field. Eqn. 7 defines the receptive field and sampling rate of the context-local block:

$$\begin{aligned} \text{RF}_{\text{ccl}} &= (r_{coarse} \times (k-1) + 1)^2, \\ \text{SR}_{\text{ccl}} &= \frac{(2 \times k^2 - 1)}{\text{RF}_{\text{ccl}}}, \end{aligned} \tag{7}$$

where r_{coarse} is the dilation rate of the coarse context branch.

A Serial of Atrous Convolutions. We define a serial of atrous convolutions: $S = ((k_1, r_1), (k_2, r_2), ..., (k_n, r_n))$. The sampling location set $\{s_i\}$ can be obtained by walking through layers. The receptive field and sampling rate are formulated in Eqn. 8.

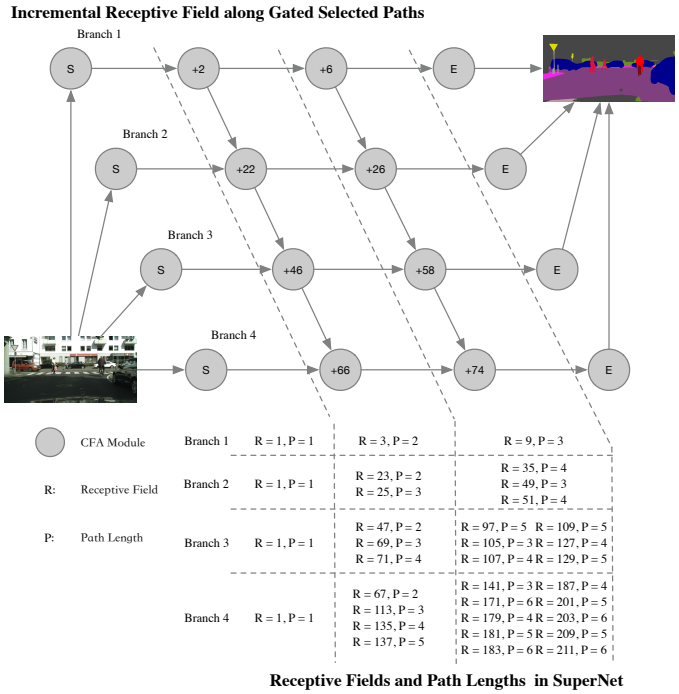


Fig. 4. The main idea of SuperNet. The incremental receptive fields along paths are filled in the grey circles in the upper part. Then internal receptive fields and path lengths are listed in the corresponding grids in the bottom part.

$$\begin{aligned} \mathrm{RF_s} &= (\sum_i (r_i \times (k_i - 1)) + 1)^2, \\ \mathrm{SR_s} &= \frac{|\{s_i\}|}{\mathrm{RF_s}}, \end{aligned} \tag{8}$$

where the operator $|\{\cdot\}|$ is used to count the number of elements of a set.

DenseASPP. Given dilation rates $\{r_1, r_2, r_3, r_4\}$, for level l in the pyramid of DenseASPP, the sample set is $P_l = \{s_i\}_l$, which is same as Eqn. 8. By overlaying all levels in the feature pyramid, the receptive field and sampling rate are:

$$\begin{aligned} \mathrm{RF_{denseaspp}} &= (\sum_i (r_i \times (k_i - 1)) + 1)^2, \\ \mathrm{SR_{denseaspp}} &= \frac{|\bigcup_l (\{P_l\})|}{\mathrm{RF_{denseaspp}}}, \end{aligned} \tag{9}$$

where the operator $\bigcup(\cdot)$ is used to calculate the union of sets.

GPSNet. Fig. 4 details the incremental receptive fields along paths in SuperNet. There are four pyramids with different ranges of scales, and each of which corresponds to a branch in GPSNet. For a specific output position, the number of sampled pixels is growing as more branches are added. Hence, the sampling location set of each branch can be defined as $P_b = \bigcup_l (\{P_{b,l}\})$. The receptive field and sampling rate of the GPS module can be calculated as follows:

$$\begin{aligned} \mathrm{RF_{gps}} &= max(RF_{b \in \{1,2,3,4\}}), \\ \mathrm{SR_{gps}} &= \frac{|\bigcup_b (\{P_b\})|}{\mathrm{RF_{gps}}}. \end{aligned} \tag{10}$$

Furthermore, Fig. 3 visualizes the receptive fields and sampling locations of GPS module. Comparing to other ASPP-like approaches, GPS module keeps dense sampling in all branches with different receptive fields.

4.2 Relation to other models

In this section, we discuss and compare *GPSNet* with the most relevant approaches including ASPP, DenseASPP, DCN, CCL and attention-based methods. With the analysis in this section, we demonstrate that *GPS* module can replace ASPP-like methods without introducing extra overhead, and other methods can also gain improvements from *GPSNet*.

ASPP. ASPP [3] adopts atrous convolution layers to segment both small and large objects. It employs multiple parallel filters with different rates to exploit multi-scale features. The extracted features from different receptive fields are further concatenated to produce the final result. Differently, by extending the parallel atrous convolution layers to grid form, the receptive field and sampling rate of GPSNet surpass ASPP by a large margin of 1.73 and 140 times respectively. Moreover, GPSNet applies soft gates to dynamically re-weight the feature maps from branches.

DenseASPP. In order to achieve large enough receptive field, DenseASPP introduces an ASPP-like module which is a cascade of atrous convolution layers. The final result is obtained from an input image that visits from small receptive fields to large receptive fields sequentially.

Instead, GPSNet introduces SuperNet with multiple entrances and exits, which is more flexible to get different scales of features. Specifically, through feeding the input in any entrances and reject it from different exits, the size of the receptive field and sampling rate are substantially improved by 0.35 and 11 times over DenseASPP respectively. Also, the inserted CFAs enable our model to generate adaptive receptive fields and sampling locations to tackle objects with large geometric deformations. Different from DenseASPP, GPS module accumulates information from large receptive fields while maintaining constant internal channels in bottlenecked branches.

DCN. To make the convolution kernel adaptive to geometric variations of the object, DCN predicts 2D offsets to augment spatial sampling locations. Our approach shares similar motivation with DCN v1 & v2 [20], [21]. But a key difference is that they focus on optimizing sampling locations on the regular grids. Although DCN is able to adjust the receptive field according to the offset, such an offset is limited by the kernel size of the corresponding standard convolution. In our method, different sampling locations in receptive fields are dynamically weighted to capture more discriminative local contexts.

CCL. CCL introduces chained context-local blocks to get multi-level contrasted local features. It produces the local representation by making a contrast between the context and local information. Actually, the context-local block can be treated as a simplified two-path ASPP, where the addition is replaced by a subtraction. Recurrent Neural Network (RNN) is further used to aggregate the multi-scale predictions. However, it needs to collect all of the features at one time, and deal with the features sequentially from high- to low-level. Comparing to the modified backbone with context-local blocks and complex modules like RNN, CFA gradually grasps different scales of information, and the enhanced feature can serve as better inputs for the subsequent prediction.

Attention-based Methods. Without considering the spatial distance, attention mechanism aggregates context via computing the similarity between features. Attention-based methods aim at capturing both short- and long-range dependencies while GPS module focuses on gathering discriminative local features. Since attention-based methods extract global context to enhance features [37], [38], those methods and GPS module can benefit each other considering the two aspects: 1) GPS module provides enhanced discriminative local features for attention to conduct the comparison which can help generate more relevant attention outputs; 2) the attention modules give global aware information to help GPS module to recognize inconspicuous objects or stuff.

GFFNet [41] proposes gated fully fusion (GFF) module to aggregate features. Here, we highlight the differences between CFA and GFF from the following three aspects: 1) *Different inputs to predict weights*. GFF module predicts gates mainly relies on the single direction features, which come from the previous layers. In comparison, CFA aims to encourage the network to aggregate sufficient but not redundant context from small to large receptive fields. In specific, our CFA module accepts features from vertical and horizontal directions, and two soft gates are generated to adaptively aggregate context information; 2) *Different ways to be integrated*. As CFA is light-weight, negligible computational overheads are introduced, making it easy to plug 7 such modules into SuperNet. However, we find that integrating GFF is nontrivial which needs to modify the whole backbone network. By contrast, our CFA module is flexible which could be readily applied to other methods; 3) *Different ranges of gate predictions*. Inspired by the prior work CCL [14], we cut off the weights with different ranges for M_v and M_h: $M_v \in [0, 1]$ is used for the smaller receptive fields, $M_h \in [-1, 1]$ is for the larger receptive fields, which make feature aggregations more flexible.

GSCNN [42]. In contrast to GPSNet which aims at adaptively aggregating contexts, GSCNN focuses on facilitating sharp predictions by highlighting boundary-related information in shape stream around object boundaries. In addition, GSCNN utilizes additional boundary supervision to train parallel shape stream, which requires to carefully modify backbones.

GPSNet. Besides receptive fields, GPSNet is the first to consider semantic segmentation from the perspective of sampling locations and the number of sampling locations simultaneously. Specifically, we introduce SuperNet, which accommodates various sub-pyramids, to maintain alterable multi-scale features. CFA is further designed to dynamically aggregate features in SuperNet. As presented in Fig. 3, GPSNet improves the size of receptive fields over ASPP-like modules like ASPP and DenseASPP. Additionally, quantitative results in Table 1 show that the sample rate significantly improves from 0.125 to 0.843, which further validates that such design choice of GPSNet is rational. Moreover, GPSNet is compute-efficient which provides a way to extract local information in complementary to existing segmentation techniques.

5 Experimental Results

GPSNet is to tackle the appearance variations and meet the contextual information demands in semantic segmentation. To demonstrate the effectiveness of our method, we conduct extensive experiments on two representative semantic segmentation benchmark datasets and compare GPSNet with previous state-of-the-art methods. In addition, complete ablation studies on baselines, ASPP and OCNet, are performed to analyze the components of GPSNet. Following the same pipeline, we plug GPS module into DenseNet [10], to further demonstrate the effectiveness of our method. The code to reproduce our results is available at https://github.com/zhaokegg/GPSNet.

5.1 Evaluation on Cityscapes

Dataset. Cityscapes is the dataset to understand urban scenes. It contains 30 common classes including road, person, car, *etc*. and only 19 of them are used for semantic segmentation evaluation. The dataset is comprised of 5,000 finely annotated images and 20,000 coarsely annotated images. The finely annotated 5,000 images are divided into 2,975, 500 and 1,525 images for training, validation, and testing. Results are evaluated with the *mean of class-wise Intersection over Union* (Mean IoU).

Training Details. On the Cityscapes dataset, we train all models with the 2,975 finely annotated images. We set the mini-batch size as 8 with InplaceABNSync [45] to synchronize the mean and standard variation. The initial learning rate is set as 0.01 and weight decay as 0.0005. Following PSPNet [6], the original image is randomly cropped to produce 769×769 input. And we employ the 'poly' learning rate policy, where the power is set to 0.9. We augment the dataset by scaling it with a factor in the rage of [0.5, 2], horizontally flipping. We train models with 40K iterations with $4\times$P40 GPUs.

Ablation Study. To investigate the effectiveness of the individual components of the proposed approach, *i.e*. SuperNet, CFA, Tuned Dilation and online hard example mining (OHEM) [47], we integrate these components into ASPP and OCNet. The ablation analysis is conducted on the Cityscapes validation set. Quantitative results are shown in Table 2 and Table 3. We report both the mIoU of OCNet as 79.58 in paper and our reproduced result as 78.70. One can observe that all of the components built on top of baseline networks consistently improve the performance. Overall, the attention-based method OCNet can benefit from GPSNet with an improvement of 0.74/1.62 (corresponding to result in paper/reproduction). It is proved that the local discriminative features provided by our method are important in parallel with global-aware information in such dense prediction task. Moreover, for a fair comparison with ASPP, we carry out experiments without OC module. Results in Table 2 show that even without the global context, the prediction of GPS module is still substantially more accurate than ASPP by 0.66. Furthermore, compared to the related method DenseASPP [7], experimental results indicate that regardless of what backbone is used, the adaptive GPS module is flexible and effective.

TABLE 2
Ablation studies on Cityscapes validation dataset. We report the results of OCNet from literature (79.58) and our reproduced experiments (78.70). In the column of multi-branch, ASP-OC means employing both ASPP and OC modules and GPS-OC means employing both GPS and OC modules.

Method	Backbone	Multi-branch	Mean IoU (%)
ASPP [3]	ResNet101	ASPP	78.65
		GPS	79.31
DenseASPP [7]	DenseNet161	DenseASPP	79.62
		GPS	79.98
OCNet [18]	ResNet101	ASP-OC	79.58 (78.70)
		GPS-OC	80.32

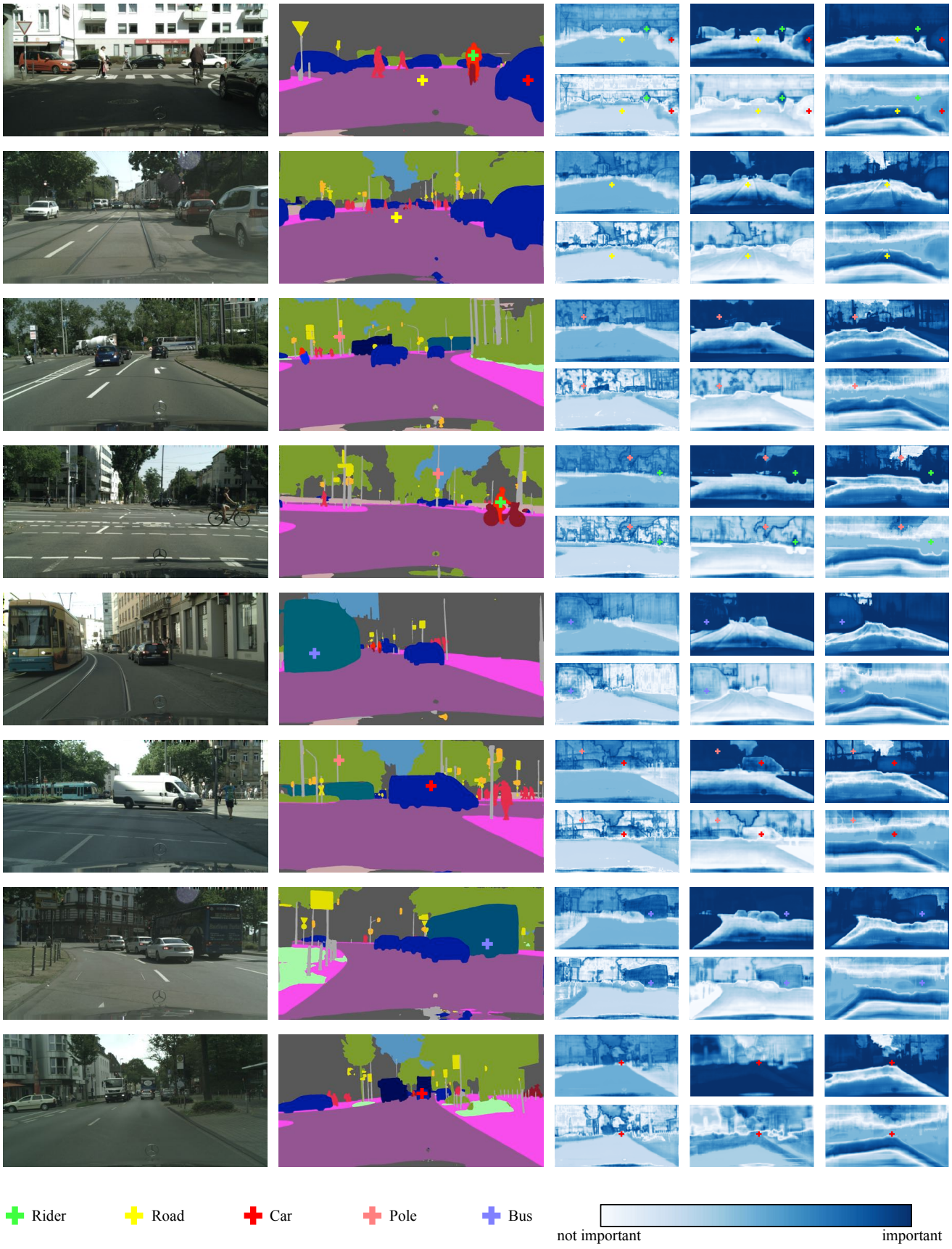


Fig. 5. Visualization of the normalized soft masks estimated by CFA on Cityscapes. The third column is the masks from g_{21}. The fourth column is the masks from g_{22}, and the last column is the masks from branch 1 and branch 4.

TABLE 3
Improvements of individual components of the proposed approach are evaluated. These ablation studies are performed on OCNet [18] to analyze the components of GPSNet.

Multi-branch	CFA	Tuned Dilation	OHEM	Mean IoU (%)
ASP	-	-	-	79.58 (78.70)
SuperNet	-	-	-	79.71
SuperNet	✓	-	-	80.03
SuperNet	✓	✓	-	80.32
SuperNet	✓	✓	✓	81.21

- **SuperNet.** To further evaluate the effectiveness of SuperNet, we compare the OCNet trained with different settings, *i.e.*, SuperNet and standard ASPP, *etc*. Validation accuracy in both settings is shown in Table 3. With SuperNet, the validation accuracy is higher than the baseline model by 0.13/1.01 (corresponding to result in paper/reproduction). It is demonstrated that SuperNet with different scales of receptive fields can help improve performance.
- **CFA** is used to adaptively select receptive fields. It can further improve the performance by 0.32 as shown in Table 3 trained with OCNet. The *SuperNet* typically benefits from the CFA where layers with larger receptive fields acquire information from the previous layers with relatively small receptive fields. CFA is able to not only control the sizes of the receptive fields but also densely select locations within the effective receptive fields. This indicates that adaptive receptive fields and sampling locations are of importance in dense object prediction to deal with object transformations.
- **Tuned Dilation.** More quantitative improvement of 0.29 with the tuned dilation is shown in Table 3. By integrating the well tuned atrous convolution layers, the network tends to gain more sampling locations that enable to densely capture semantic contexts. The result further shows that tuned dilation is indeed effective for increasing sampling rates to improve the performance.
- **OHEM.** To tackle with data imbalance and overfitting, we further conduct experiments to validate the effectiveness of OHEM as shown in Table 3. Following previous work [47], we set the threshold for selecting hard pixels as 0.7, and keep at least 100,000 pixels within each mini-batch. The result shows that the OHEM built on our network can further boost the performance.

TABLE 4
Comparison of computational costs. We calculate FLOPs of different methods excluding the backbone, where the input image size is 2048×1024.

Method	#FLOPs.(G)
ASPP	1236.9
DenseASPP	911.6
GFFNet	750.8
GSCNN	1386.9
SuperNet	249.0
GPSNet	331.8

Performance. On Cityscapes, we compare GPSNet with several competitive baselines including the dilation-based methods, *i.e.*, DeepLabv3 [4], DUC-HDC [15], DenseASPP [7], region-based method *i.e.*, PSPNet [6], and attention-based method *i.e.*, PSANet [16], OCNet [18]. We evaluate our results on the Cityscapes testing set with multi-scale testing. The results are shown in Table 5. The prediction of GPSNet is substantially more accurate than the methods conducted with ResNet-101. Notably, our result also outperforms DenseASPP which takes DenseNet-161 as the backbone. In addition, we also measure floating point operations (FLOPs) for different methods in Table 4. From the results, we observe that GPSNet is $2\times$ more compute-efficient than other approaches. Visual results are shown in Fig. 5.

TABLE 5
Results on Cityscapes test dataset. The results marked with † indicate the models trained without validation set. The result of Deeplabv3 marked with $*$ is trained with both finely and coarsely annotated training data.

Method	BaseNet	Mean IoU (%)
DenseASPP [7]	DenseNet161	80.6
Deeplabv3* [5]	ResNet101	81.3
DUC-HDC [15]	ResNet101	77.6
PSPNet [6]	ResNet101	78.4
PSANet [16]	ResNet101	78.6
OCNet† [18]	ResNet101	80.1
OCNet	ResNet101	81.2
GPSNet†	ResNet101	80.6
GPSNet	ResNet101	**82.1**

5.2 Evaluation on ADE20K

Dataset. The scene parsing dataset ADE20K contains 150 classes and diverse complex scenes with 1,038 image-level categories. It needs to parse both objects and stuff. The dataset is divided into 20,000, 2,000 and 3,000 for training, validation and testing. Results are evaluated with Mean IoU.

Training Details. On the ADE20K dataset, the base learning rate is set as 0.02 and with a weight decay 0.0001. The input image is resized to 480. The mini-batch size is 16 and we also apply InplaceABNSync to synchronize the mean and standard deviation across multiple GPUs. The models are trained with 200K iterations with $4\times$P40 GPUs. The learning rate policy and data augmentation are the same as those on the Cityscapes dataset.

Performance. On ADE20K, we compare our evaluated *GPSNet* with three attention-based method, *i.e.*, PSANet [16], EncNet [17], OCNet [18], gated feature fusion network (GFFNet) [41], and region-based method, *i.e.*, PSPNet. The experiments are evaluated on the ADE20K validation set. The results reported in Table 6 show that *GPSNet* consistently outperforms all baselines. Notable, *GPSNet* surpasses the deeper 269-layer PSPNet by 0.82.

5.3 Understanding GPSNet

To illustrate the ability of dynamically extracting discriminative local features, we visualize the soft masks produced by CFA and context aggregation predicted by GPS module.

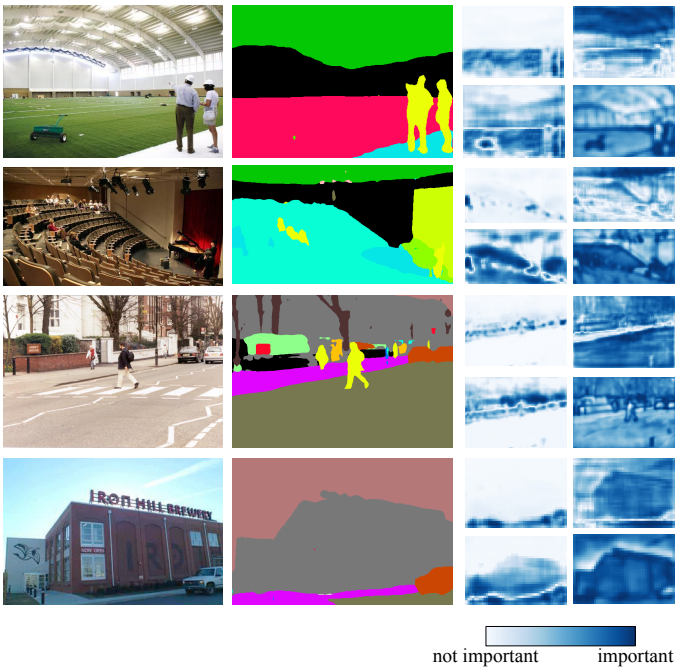


Fig. 6. Visualization of the normalized soft masks estimated by CFA on ADE20K. The third column is the masks from g_{31}. The fourth column is the masks from branch 3 and branch 4.

TABLE 6
Results on ADE20K validation dataset.

Method	BaseNet	Mean IoU (%)
PSPNet [6]	ResNet269	44.94
PSPNet	ResNet101	43.29
PSANet [16]	ResNet101	43.77
EncNet [17]	ResNet101	44.65
OCNet [18]	ResNet101	45.45
GFFNet [41]	ResNet101	45.33
GPSNet	ResNet101	**45.76**

5.3.1 CFA Visualization

Fig. 5 depicts the masks produced by CFA on Cityscapes. As Cityscapes focuses on understanding automatic driving scenes, we highlight the pixels from relative hard objects, *e.g.*, rider, road, car, pole and bus.

In the automatic driving scenes, CFA can adaptively select the receptive fields and sampling locations to capture the local contexts and we have the following observations:

- For movable objects like car and rider, the main features are captured with branch 1, in which the receptive field is the smallest.
- For texture-less regions like road, the features are captured with branch 4, in which the receptive field is the largest.
- To aggregate boundary features, the features from the convolution layer with dilation rate r_{11} are assigned with large weights.
- For large objects like buses, the features from branch 4 are given larger weights than small objects.
- For tiny objects like poles, the local features from branch 1 are much more important than those from branch 4.

In addition, we also visualize the masks from ADE20K [24] in Fig. 6. Compared with street view images from Cityscapes, web photos are with lower resolution and more classes. The dataset emphasizes more on the target context which takes a great proportion of the whole image. In fact, the limited resolution and dominated objects force CFA to predict much more discriminative local features. Followings are the observations from our experiments:

- For tiny indoor areas, which are not salient in the surrounding context, corresponding receptive fields are also shrunk to avoid misclassification.
- To classify dominated foreground like building, receptive fields are expanded to collect more context.
- Because of the limited resolution, the features from branch 3 are assigned with large weights to classify regions like road, which slightly differ from the above observations but keep a consistent trend.

As shown in the above observations, the soft masks predicted by CFAs are relative with scene, class and scale. To make features discriminative, each inserted CFA selects more relevant contexts from larger receptive fields. On the one hand, it validates the conclusion of existing works that enlarged receptive fields can improve recognition. On the other hand, our experiments demonstrate that CFA further improves recognition accuracy with adaptive receptive fields and sampling locations. Even in different scenes, CFA is able to adaptively select discriminative context based on the inputs.

5.3.2 GPS Module Visualization

With gradually inserted CFAs, GPS module grasps informative discriminative local features by weighting different sampling locations in growing receptive fields. To verify the effectiveness of different sampling locations, we approximately use the production of predicted weights on the paths to visualize the free form receptive fields. As shown in Fig. 7, the relevant contexts associated with given locations are highlighted. The mechanisms of learned GPS module to improve segmentation results are analyzed as follows:

- In GPS module, the maximum receptive fields are large enough to cover scale ranges of objects in the scene. Furthermore, locations around the target pixels are densely sampled.
- To keep more relevant contexts, several locations in the receptive fields are given small weights.
- GPS module is likely to predict large weights for the target locations. We can find that the main impact to extract semantic context is from the locations inside the objects, which is in line with [48].
- The surrounding compatibility contexts like road or rider, are also aggregated as complements to eliminate confusing ambiguity. Contextual information is important for locations that lack sufficient discrimination.

The visualization results validate that learned GPS module is able to adaptively aggregate reasonable contexts. By analyzing the data-dependent receptive fields and sampling locations, we can observe that the learned mechanism is consistent with the existing works. However, our GPS module can be trained in an end-to-end manner and achieve higher performance.

6 Conclusion

In this paper, we have presented the *Gated Path Selection Network (GPSNet)* to optimize adaptive receptive fields, sampling

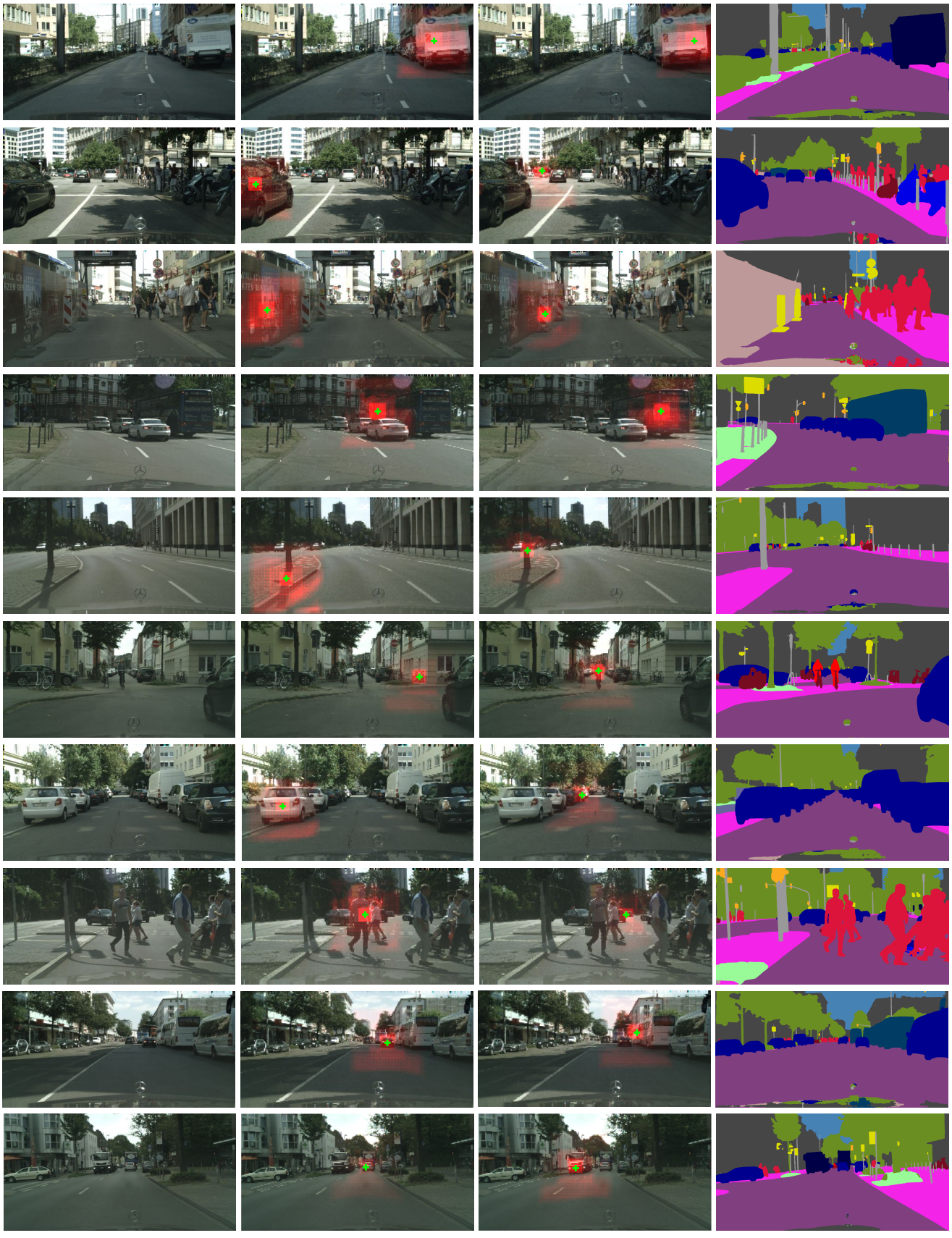


Fig. 7. Visualization of our segmentation results on Cityscapes and corresponding sampling locations predicted by the GPS module. In column 2 and column 3, the selected locations are marked with green dots, and the regions with dark red are more relevant to the corresponding locations.

locations and the number of samples simultaneously in semantic segmentation. A *SuperNet* is proposed to ensure maximum information flow in the network. It provides various paths to extract multi-scale representations. Dense connectivity in the network allows the network to effectively aggregate contexts from larger receptive fields. The strategy to adjust the dilation rates in the atrous convolution grid makes the sampling locations much denser. *Comparative Feature Aggregation* is further introduced to estimate soft masks to dynamically select effective context locations and regularize the receptive fields. Besides, our method is simple, efficient and model-agnostic. It can be applied to various ASPP-like architectures. The proposed method has shown its effectiveness on two competitive semantic segmentation datasets, *i.e.*, Cityscapes, ADE20K, and achieves new state-of-the-art results. Future research may focus on extending our results to other types of computer vision tasks, such as object detection and image generation.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Develop Program of China under Grant No.2018YFB2100601, and National Natural Science Foundation of China (NSFC) under Grant No. 61872024.

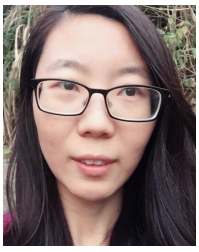
REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
[2] 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.
[3] ——, "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, 2017.
[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.
[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
[7] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *CVPR*, 2018.
[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.
[11] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *CVPR*, 2017, pp. 4353–4361.
[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *ECCV*, 2018.
[13] 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*. Springer, 2015, pp. 234–241.
[14] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *CVPR*, 2018.
[15] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *WACV*. IEEE, 2018.
[16] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018.
[17] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *CVPR*, 2018.
[18] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.
[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *CVPR*, 2019.
[20] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017.
[21] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *CVPR*, 2019.
[22] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *CVPR*, 2017, pp. 4201–4209.
[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
[24] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.
[25] 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 *ICCV*, 2015.
[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[27] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.
[28] 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.
[29] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.
[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[31] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.
[32] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 603–619.
[33] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *ICLR Workshop*, 2016.
[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *ICCV*, 2019.
[35] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.
[36] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *ICCV*, 2019, pp. 9167–9176.
[37] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *ICCV*, 2019.
[38] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
[39] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 82–92.
[40] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8553–8562.
[41] X. Li, H. Zhao, L. Han, Y. Tong, and K. Yang, "Gff: Gated fully fusion for semantic segmentation," *arXiv preprint arXiv:1904.01803*, 2019.
[42] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5229–5238.
[43] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *arXiv preprint arXiv:1808.03833*, 2018.
[44] 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.

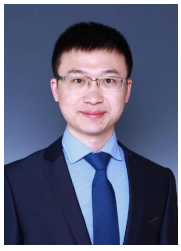
[45] S. Rota Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *CVPR*, 2018.
[46] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, and K. Weinberger, "Convolutional networks with dense connectivity," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
[47] Z. Wu, C. Shen, and A. v. d. Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *arXiv preprint arXiv:1604.04339*, 2016.
[48] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in neural information processing systems*, 2016, pp. 4898–4906.

Qichuan Geng is a Ph.D. candidate, at State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He received his B.S. degree from Beihang University in 2012. His main research interests include computer vision, semantic segmentation and scene geometry recovery.

Hong Zhang is currently a senior software engineer at Baidu Research. She obtained her MPhil degree from The Chinese Univerisity of Hong Kong and B.S. degree from Wuhan University. Her research areas include computer vision and machine learning.



Xiaojuan Qi received her B.Eng degree in Electronic Science and Technology at Shanghai Jiao Tong University (SJTU) in 2014, and the PhD degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2018. She was a postdoctoral researcher at the University of Oxford in 2019. She is an assistant professor at the University of Hong Kong. Her research interests include artificial intelligence, computer vision, robotics, deep learning, and medical image analysis.



Gao Huang is an Assistant Professor in the Department of Automation at Tsinghua University. Before joining Tsinghua University, he was a Postdoctoral Researcher in the Department of Computer Science at Cornell University. He received the PhD degree in Control Science and Engineering from Tsinghua University in 2015, and B.Eng degree in Automation from Beihang University in 2009. His work on DenseNet won the Best Paper Award of CVPR in 2017. His research interests include deep learning and computer vision.

Ruigang Yang is currently CTO of Inceptio, on leave from the University of Kentucky where he is a full professor. He was the Chief Scientist for 3D Vision at Baidu Research, where he established and lead the Robotics and Autonomous Driving Lab (RAL). He obtained his PhD degree from University of North Carolina at Chapel Hill and his MS degree from Columbia University. His research interests span over computer vision and computer graphics, in particular in 3D perception and 3D synthesis. He has published over 100 papers, which, according to Google Scholar, has received over 10000 citations with an h-index of 58 (as of 2020). He has received a number of awards, including US NSF Career award in 2004, best demonstration award in CVPR 2006, and University of Kentucky's Dean's Research Award in 2013. He is currently a senior member of IEEE.

Zhong Zhou Professor, Ph.D adviser, State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He got B.S. degree from Nanjing University and Ph.D. degree from Beihang University in 1999 and 2005 respectively. His main research interests include Virtual Reality/Augmented Reality/Mixed Reality, Computer Vision and Artificial Intelligence. He is member of IEEE, ACM and CCF.