Gaussian Dynamic Convolution for Efficient Single-Image Segmentation

Xin Sun, *Member, IEEE*, Changrui Chen, Xiaorui Wang, Junyu Dong, *Member, IEEE*, Huiyu Zhou, and Sheng Chen, *Fellow, IEEE*

***Abstract*— Interactive single-image segmentation is ubiquitous in the scientific and commercial imaging software. Lightweight neural network is one practical and effective way to accomplish the single-image segmentation task. This work focuses on the single-image segmentation problem only with some seeds such as scribbles. Inspired by the dynamic receptive field in the human being's visual system, we propose the Gaussian dynamic convolution (GDC) to fast and efficiently aggregate the contextual information for neural networks. The core idea is randomly selecting the spatial sampling area according to the Gaussian distribution offsets. Our GDC can be easily used as a module to build lightweight or complex segmentation networks. We adopt the proposed GDC to address the typical single-image segmentation tasks. Furthermore, we also build a Gaussian dynamic pyramid Pooling to show its potential and generality in common semantic segmentation. Experiments demonstrate that the GDC outperforms other existing convolutions on three benchmark segmentation datasets including Pascal-Context, Pascal-VOC 2012, and Cityscapes. Additional experiments are also conducted to illustrate that the GDC can produce richer and more vivid features compared with other convolutions. In general, our GDC is conducive to the convolutional neural networks to form an overall impression of the image.**

***Index Terms*— Image segmentation, convolutional neural networks, weakly supervised learning, dynamic receptive field.**

I. Introduction

SEMANTIC segmentation aims to compute a dense label prediction for each pixel in an image. It is used ubiquitously across all scientific and commercial fields where imaging has become the most critical step [1]. Recent success of semantic segmentation lies on the end-to-end training with large-scale segmentation annotations. In scientific and commercial software, however, interactive image segmentation from a single image is recognized as a most user-friendly and practical operation. For example, Quick Selection tool in Adobe Photoshop is a typical commercial implementation of scribble-based interactive segmentation method. It does not rely on a large training dataset. Instead, it needs to be optimized on each image independently. Such single-image segmentation task is vital in interactive segmentation [2]–[4] and weakly supervised segmentation. This work proposes the Gaussian dynamic convolution (GDC) to build a lightweight convolution neural network for fast and effectively accomplishing the single-image segmentation task. The network can be optimized in a flash only with one image and some seeds such as the scribbles. The proposed GDC can be easily integrated into various convolution modules. The background and motivation of our proposed GDC are now further elaborated.

The essence of semantic segmentation is to identify distinctive features for different categories. The receptive field is of great significance for the segmentation task [5], [6]. Deep features with different receptive fields represent various levels of the visual attributes [5]. It is informative to first revisit the way that human beings observe the visual world. Undoubtedly, there is a receptive field mechanism in the human visual system. The question is that *do human beings adopt a fixed scale or a group of fixed scales of receptive fields when they observe objects?* The receptive field in the human being's visual system is totally dynamic [7]. During the long period of growing up from a baby to an adult, the dynamic receptive field provides us stereoscopic and vivid information about the world [8]. What we store in our minds are living objects rather than some rigid features in several scales. Therefore, it is far better to equip the neural networks with stochastic dynamical receptive field for flexibly capturing context. To this end, we propose the GDC, a novel convolution kernel with a dynamic receptive field, to extract richer features for segmentation tasks.

Researchers have already realized that the multi-scale features and the large receptive field [5], [6] are profitable for segmentation tasks. Usually a pyramid architecture is adopted with some dilated convolution or some pooling operation to obtain the multi-scale features with a large receptive field.

Manuscript received April 17, 2021; revised June 25, 2021; accepted July 8, 2021. Date of publication July 13, 2021; date of current version May 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Project 61971388, Project U1706218, and Project L1824025; and in part by the European Union's Horizon 2020 Research and Innovation Program through the Marie-Sklodowska-Curie Grant under Grant 720325. This article was recommended by Associate Editor J. Lu. *(Corresponding author: Xin Sun.)*

Xin Sun is with the Department of Computer Science and Technology, Ocean University of China, Qingdao, Shandong 266100, China, and also with the Department of Aerospace and Geodesy, Technical University of Munich, 80333 München, Germany (e-mail: sunxin1984@ieee.org).

Changrui Chen is with the WMG Data Science, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: geoffreychen777@gmail.com).

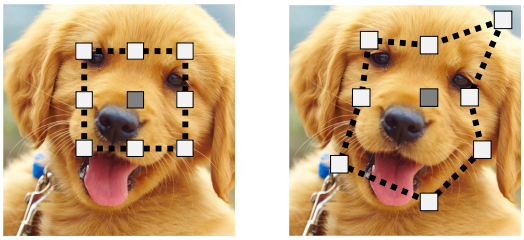
Xiaorui Wang and Junyu Dong are with the Department of Computer Science and Technology, Ocean University of China, Qingdao, Shandong 266100, China (e-mail: recyclerblacat@stu.ouc.edu.cn; dongjunyu@ouc.edu.cn).

Huiyu Zhou is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: hz143@leicester.ac.uk).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: sqc@ecs.soton.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3096814.

Digital Object Identifier 10.1109/TCSVT.2021.3096814



(a) (b)

Fig. 1. (a) The fixed receptive field of the dilated convolution, and (b) the dynamic receptive field of the proposed GDC.

Fig. 1a illustrates the receptive field of the dilated convolution using a toy example. Dilated convolution can expand the receptive field but it can only supply the same scale features. By contrast, our GDC is capable of overcoming the limitation of the dilation factors. The receptive field of the GDC is illustrated in Fig. 1b, where the convolution kernel dynamically selects convolutional positions, and the weight vectors are scattered to different positions to extract features in diverse scales. Consequently, the dynamic convolution kernel can not only extend the receptive field but also produce richer feature maps with variable receptive fields.

More specifically, the GDC stochastically forms the convolution kernel of various spatial scales according to some Gaussian distribution offsets. Owing to this randomness, the GDC can supply more diverse feature maps to the following networks. We first employ the GDC to build a lightweight segmentation network for the single image segmentation. Furthermore, we also introduce the GDC to build a Gaussian dynamic pyramid pooling (GDPP) module for the traditional semantic segmentation task, in order to show its generality. The experiments conducted demonstrate that the proposed GDC achieves competitive object semantic segmentation results on the Pascal-Context [9], Pascal-VOC 2012 [10], and Cityscapes datasets [11]. We also conduct some explanatory experiments to discuss why the proposed GDC works.

In summary, the main contributions of this paper are:

- We propose a novel GDC with a dynamic receptive field to aggregate richer features in various scales.
- Our GDC can be easily and efficiently inserted into various convolutional modules for better segmentation performance.
- Our experiments indicate that the GDC achieves the state-of-the-art performance in the single image segmentation and semantic segmentation tasks on the three datasets, Pascal-Context, Pascal-VOC 2012, and Cityscapes.
- We explain the mechanism and properties of the proposed GDC by additional experiments.

The rest of this paper is organized as follows. We briefly introduce the related work in Section II. Our GDC is proposed in Section III. Section IV introduces the two typical segmentation networks with our GDC. The experiments are conducted in Section V. In Section VI, we analyze the mechanism of this novel GDC and discuss why it works. Finally, our conclusions are offered in Section VII.

II. Related Work

Semantic segmentation is one of the fundamental tasks in computer vision and it benefits a variety of applications, ranging from biometric identification to object recognition [12]. The essence of semantic segmentation is to identify distinctive features for different categories. Researchers in the early years tried to extract the handcrafted features, such as the color feature or the geometric feature, to discriminate the labels of all the pixels. Thanks to the resurgence of deep convolutional neural networks (DCNNs), segmentation technology has made great progress in the past few years. Driven by powerful DCNNs [13]–[17], the deep segmentation networks, such as FCN [18], SegNet [19], PSPNet [6], and DeepLabs [20], [21], achieve the state-of-the-art performance of the semantic segmentation task on different benchmark datasets. In particular, FCN [18] adopts a fully convolutional network and is optimized by end-to-end training. SegNet [19] shares the same idea to design an encoder-decoder architecture and uses some skip connections to utilize the low-level features. For more fine tasks, skip connections are not enough to help accurately locating indistinct boundaries. Zhou *et al.* [22] proposed a novel high-resolution multi-scale encoder-decoder network, in which multi-scale dense connections are introduced for the encoder-decoder structure to exploit comprehensive semantic information. SSAP [23] uses the pixel-pair affinity pyramid, combines the affinity of pixel pairs and semantic segmentation at different scales, and improves the predictions of instances level by level starting from the deepest layer. Wang *et al.* [24] focus on multi-level features to object segmentation. Their conditional Boltzmann machine is suitable to map multi-level convolutional features of object parts onto the global shape of object. Furthermore, context representations have been widely used to profit semantic image segmentation. SCN [25] novelly uses the local structural feature maps to compute the context representations in top-down switchable information propagation. The context representations are combined with the convolutional features to form the intermediate feature maps, which are used for the final semantic segmentation. Ji *et al.* [26] proposed locality-preserving CNN, which uses relationship between similar samples to auxiliary segmentation. Experiments show that locality-preserving is more suitable for small sample segmentation process. Lin *et al.* [27] proposed cross domain complex learning, which effectively utilizes the segmentation labels of synthetic images and variation of real images through an auxiliary task. This method has been proved to be able to transfer context information knowledge from domain to domain in some specific tasks

In contrast to the semantic segmentation, single image segmentation does not rely on a training dataset. Instead, the segmenting method needs to be optimized on each image independently. This task is vital in interactive segmentation [2]–[4] and weakly supervised segmentation [28], [29]. In recent years, many researches [30]–[36] focus on weakly supervised segmentation to conquer the problem of scarcity of labeled data. One of the mainstream strategies for this task is to transform the weak labels, such as points [37], [38], scribble [39], [40], bounding box [35], [41] or image-level label [42], to a coarse segmentation ground truth for the fully

supervised segmentation networks. Hong *et al.* [43] provide a comprehensive overview of weakly supervised approaches for semantic segmentation. Specifically, they point out the limitations of various supervision level methods and discuss the directions worthy of effort to improve performance.

For the weakly supervised segmentation based on image-level labels, STC [42] adopts a simple-to-complex framework and proposes a method for obtaining refined segmentation results with three progressively powerful DCNNs. Redondo-Cabrera *et al.* [44] propose a hide-and-seek strategy which consists of two class activation mapping (CAM) modules so as to recover the activation masks covering the full object extents by randomly hiding patches in a training image, forcing the second CAM network to seek other relevant parts. Chen *et al.* [45] utilize a self-supervised scheme without any ground truth to promote the saliency detection and image segmentation results. Meng *et al.* [46] Proposed a new segmentation strategy, which first segmented the foreground by class-level, and then fused all the foreground information to get the final result. The weaklier supervised semantic segmentation [47] imposes even more challenges, as it uses only one image-level annotation per category to achieve a desired semantic segmentation performance. Furthermore, for each category, one sample has image level annotation, while only the number of object categories contained in each image is provided for other samples. Mixed-use of image-level labels and bounding box labels can further improve performance [48].

For scribble supervision, ScribbleSup [39] uses some scribbles to generate the segmentation results for an image. GraphNet [49] combines the deep network with the graph structure. These methods use the scribble annotations to generate the pseudo annotations by the graph convolution. However, ScribbleSup generates segmentation proposals from scribbles and uses these proposals to alternatively train an FCN, which can easily be trapped in local minimums. To solve this problem, Tang *et al.* [40] abandon the alternating training method and train a FCN via a joint loss function with two terms: the partial cross-entropy loss for scribbles only and the relaxed normalized-cut regularizer that implicitly propagates the true labels to unknown pixels during training. Shen *et al.* [50] jointly train the weakly supervised object detection and segmentation tasks to complement each other's learning. Such a cross task enforcement helps both the tasks to leap out of their respective local minimums. The work [51] presents a boundary perception guidance (BPG) approach that only leverages scribbles. It consists of two basic components which are prediction refinement and boundary regression to make better segmentation progressively. Wang *et al.* [52] incorporate convolutional neural networks (CNNs) into a bounding box and scribble-based binary segmentation pipeline to resolve the problem of interactive 2D and 3D medical image segmentation. Ji *et al.* [53] present a scribble-based hierarchical weakly supervised learning pipeline for medical image structure segmentation which integrates graph-based method with only whole tumor/normal brain scribbles and the global labels. This work was the first to realize such a weak supervision level in the field of compression structure segmentation. Lu *et al.* [54] introduce the Boundarymix method which generates pseudo-training images for improving segmentation with scribble annotations. Moreover, Zhang *et al.* [55] propose to use scribble annotations for weakly-supervised salient object detection.

There is a contradiction in deep networks for segmentation. Receptive field plays a critical role in the segmentation task. Conventional deep convolution networks expand the receptive field through stacking plenty of convolution layers and pooling layers. This strategy exposes a defect. There is a huge resolution disparity between the feature maps and the input images. Many variations of the normal convolution attempt to resolve this paradox. Dilated convolution [5] inserts some gaps into the convolution kernel to support an exponential expansion of the receptive field. Deformable convolution [56], which is commonly used for detection, adopts a side network to learn an offset for each weight vector in the convolution kernel. Both these two approaches can effectively expand the receptive field without losing the resolution of the feature maps. Another essential factor for segmentation is the multi-scale features. Researches such as PSPNet [6] and DeepLabs [21] endeavor to extract the multi-scale features through a pyramid architecture. They use different dilated factors or different pooling strides in different layers of the pyramid to capture multi-scale features.

Almost all the existing methods of image segmentation discussed in this section use normal or dilated convolution. Different from dilated convolution or deformable convolution, our GDC is simpler and more effective in many cases. For example, the deformable convolution requires several extra layers of a small network to learn the offset. In contrast, our GDC dynamically changes the offsets of the weight vectors in the convolution kernel. Owing to its flexible dynamic range of receptive fields, our GDC can generate rich features. In this paper, we use our GDC to build a lightweight CNN to accomplish the labels transformation. Our lightweight segmentation network only requires one image with some seeds such as the scribbles, and it can be optimized in a flash. Since our GDC introduces dynamics and random factors, it outperforms the existing convolution methods for the one-image segmentation with some seeds such as the scribbles. Moreover, our GDPP built on the proposed GDC offers a competitive means to the existing state-of-the-arts method for the semantic segmentation.

III. Gaussian Dynamic Convolution

Without loss of generality, we use the 3×3 convolution kernel as an example to illustrate the proposed GDC, as shown in Fig. 2.

To highlight the difference of our GDC with other types of convolution, we start from the normal convolution kernel, where a regular 3×3 grid slides over the feature maps F^l at layer l to sample 9 feature vectors. A normal convolution kernel summarizes the sampled vectors weighted by the corresponding weights in the convolution kernel. Let $c = <i, j>$ be the coordinate of the center feature vector in one convolution operation. Then, the coordinate $\overline{c}$ of the other eight sampled feature vectors can be calculated with a direction basis e and

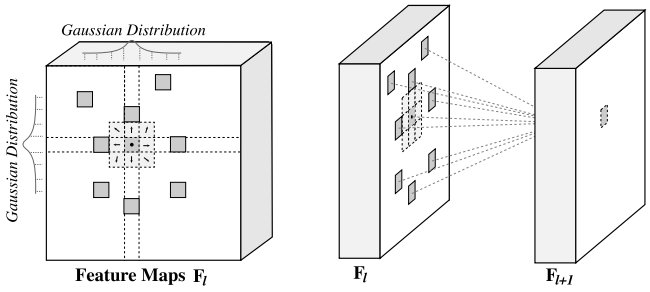


Fig. 2. Illustration of the GDC with kernel size 3.

an offset value Δ according to

$$\overline{c}_i = c + \Delta_i \odot e_i,$$
$$e_i \in \mathcal{E} = \{< -1, -1 >, < -1, 0 >, \cdots, < 1, 1 >\}, \quad (1)$$

where the direction basis e denotes the basic sample direction, the offset value Δ is always $< 1, 1 >$ in a normal 3×3 convolution kernel, and $\odot$ denotes the element-wise product operator. The element-wise product may be replaced by the Hadamard product. For example, the coordinate of the sampled feature vector at the left-top corner can be calculated by:

$$< i, j > + < 1, 1 > \odot < -1, -1 > = < i - 1, j - 1 > . \quad (2)$$

For the dilated convolution [5], we can also calculate the coordinates of the sampled feature vectors by Eq. (1). The dilated factor is equal to the offset value Δ. The offsets may make the location not in a regular grid, which is a common problem in image processing. And we use the simple down rounding to save the calculation time and memory.

The proposed GDC is illustrated in Fig. 2. For each convolution position, we fix the center weight of the convolution kernel. Then we scatter the other convolution weight vectors by randomly changing the offset value Δ. More specifically, we sample the offset values from a two-dimensional half Gaussian distribution with standard deviation Σ, denoted by $HalfGaussian_2(0, \Sigma)$. The one-dimensional half Gaussian distribution is expressed as:

$$HalfGaussian(0, \Sigma) = \frac{\sqrt{2}}{\Sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\Sigma^2}\right), \quad x \geq 0. \quad (3)$$

Suppose that c is the coordinate of the center convolution position. The coordinates $\tilde{c}$ of the other convolution positions in the GDC can also be calculated by Eq. (1). But instead of using a fixed offset value $\Delta = < 1, 1 >$, Δ in our GDC obeys $HalfGaussian_2(0, \Sigma)$. Different Δ values will produce different feature maps. The summation operation of the sampled feature vectors in the GDC is the same as in the normal convolution.

IV. Gaussian Dynamic Convolution Based Image Segmentation Networks

A. Single-Image Segmentation

Single image segmentation is quite different from the traditional semantic segmentation. Usually, it needs to optimize the algorithm and generates the segmentation result all based on only one image. For the single image segmentation task, DCNNs are inapplicable. This is because conventional deep segmentation networks are typically composed of plenty of convolutional layers to obtain more information. These DCNNs need to be trained with a large set of data. One image is insufficient to support the training of a deep network, and the optimization time of a deep network is too high. Ideally, a lightweight network is preferred, which can be optimized extremely fast and can generate satisfactory segmentation result. In this subsection, we address the challenges imposed by the single image segmentation task, and design a lightweight segmentation network with the novel GDC proposed in the previous section. This lightweight network based on the GDC can be optimized in a flash and it is capable of generating satisfactory segmentation result.

1) Network Architecture: As shown in Fig. 3, we use the first layer and the first convolutional linear bottleneck of the MobileNet v2 [57] as our feature extractor. These two network components generate two groups of feature maps with only 16 and 24 channels, respectively. We resize these two groups of feature maps to the half size of the input image by the bilinear interpolation. Then, the feature maps are concatenated and sent into a 1×1 convolutional layer to fuse the channel information. After that, we send the fused feature maps to the GDC module.

The GDC Module consists of two branches. The first one is a normal 3×3 convolution layer which is used to fuse the local features. The second branch is our GDC. We use this branch to gather features in various scales.

Finally, we send the feature maps generated by the GDC module to a pixel-level classifier to estimate the final segmentation result. The pixel-level classifier is composed of a 3×3 convolution layer followed by a ReLU activate function, and a 1×1 convolution layer with a softmax layer.

2) Training: We initialize and optimize our lightweight segmentation network for each image, instead of using plenty of training data. In our experiment, the training ground truth is generated by some weakly semantic cues such as the scribbles [39]. This segmentation network is extremely fast as it consists of very few convolution layers. The detailed training of this lightweight GDC segmentation network is given in Section V.

B. Semantic Segmentation

As discussed previously, multi-scale features are essential for achieving good performance in semantic segmentation tasks. Traditionally a pyramid architecture is adopted with dilated convolution to obtain the multi-scale features with a large receptive field. In order to show the potential and generality of the proposed GDC, we further design a variant GDC as shown in Fig. 4 suitable for the application to semantic segmentation. Specifically, we scatter the weight vectors of the convolution kernel by a same offset value Δ. The offset values Δ are sampled from a two-dimensional half Gaussian distribution with standard deviation Σ as shown in Fig. 4. We set a constant base offset and calculate the sampled position by introduce a hyperparameter Δ_{base} in the

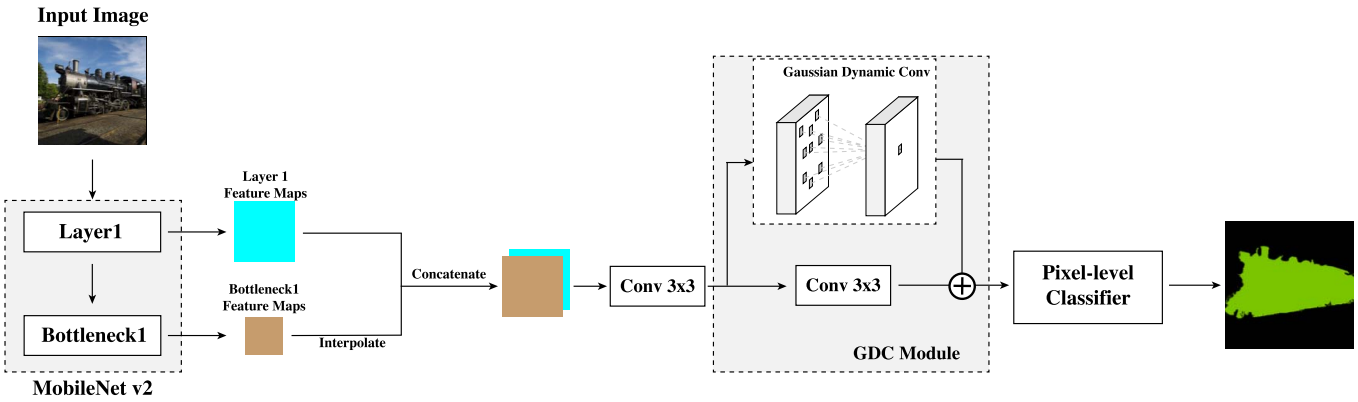


Fig. 3. Pipeline of the lightweight segmentation network with the GDC. We adopt the first layer and the first convolution linear bottleneck of the MobileNet v2 to extract two groups of feature maps.

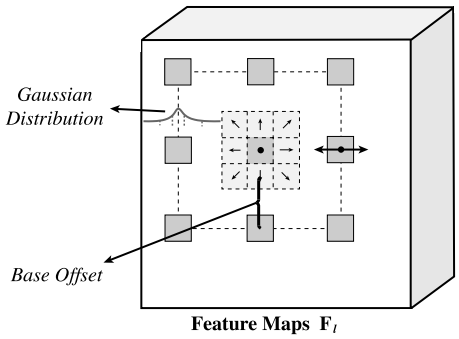


Fig. 4. Illustration of the variant GDC with kernel size 3.

variant GDC. The sampled position can be calculated by:

$$\overline{c}_i = c + (\Delta_{base} + \Delta) \odot e_i, \tag{4}$$

where c is the coordinate of the center convolution position.

It is clear that this variant GDC degenerates into the dilated convolution when the dynamic offset $\Delta = 0$. The constant base offset or hyperparameter Δ_{base} ensures that our dynamic convolution can expand the receptive field as the dilated convolution does. Moreover, the dynamic offset Δ brings richer features for the following network components. We adopt this variant GDC to build a GDPP for implementing the semantic segmentation network.

Multi-scale feature fusing modules for semantic segmentation usually use several dilated convolution layers with different factors to gather multi-scale information. In Fig. 5, the small yellow square denotes the dilated convolution layer with a small factor which can be used to gather the small scale information. On the other hand, the dilated convolution denoted by the large red square is used to obtain large scale information. To fuse the middle-scale information, other modules, such as ASPP [21], use one or more dilated convolutions with different fixed middle factors. Our GDPP module is different in that we use the GDCs to gather the information to implement various middle scales, as shown in the bottom part of Fig. 5. In other words, in our GDPP module for semantic segmentation depicted in Fig. 5, the largest and smallest

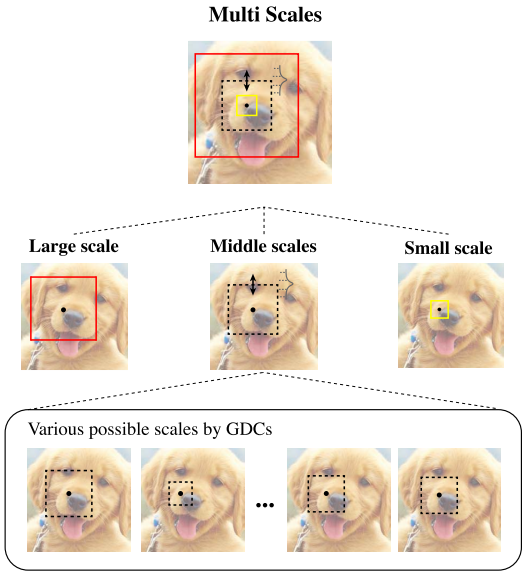


Fig. 5. Illustration of the Gaussian dynamic pyramid pooling, where the largest and smallest scales are fixed, while the middle scales are dynamically sampled by the Gaussian distribution.

scales are fixed to limit the range of GDCs, while the middle scales are dynamically sampled by the Gaussian distribution. Therefore, the proposed GDPP module can produce richer and more vivid feature maps during the training phase.

V. Experiments

The experiments (both training and testing) were conducted on a PC with a GTX 2080ti GPU. The implementation code will be published on https://github.com/ouc-ocean-group/.

A. Single Image Segmentation

1) Implementation: Single image segmentation plays a vital role in many other tasks including weakly supervised segmentation label generation. The sizes of the images in single image segmentation are variable. Therefore, we adopt an adaptive Gaussian dynamic offset of $s \times \Delta$, where s is the length of the

TABLE I
PERFORMANCE COMPARISON OF THE SINGLE IMAGE SEGMENTATION ON TWO DIFFERENT DATASETS OF PASCAL CONTEXT AND PASCAL-VOC 2012

	Pascal Context		Pascal-VOC 2012 (val)	
Methods	**Overall Acc.**	**mIoU (%)**	**Overall Acc.**	**mIoU (%)**
Normal 3x3 conv	0.7900	53.41	0.8887	68.94
Dilated conv (6)	0.8327	60.01	0.9047	71.95
Dilated conv (16)	0.8324	59.86	0.9034	71.80
Dilated conv (24)	0.8308	59.57	0.9029	71.78
Deformable conv	0.8003	54.67	0.8883	68.69
GDC ($\Sigma = 0.2$)	**0.8646**	**65.12** ▲11.71	**0.9092**	**74.06** ▲5.12

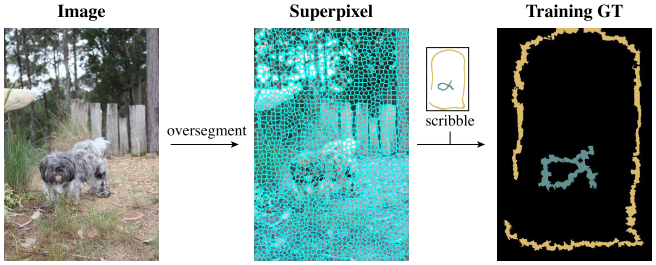


Fig. 6. Illustration of the training ground truth expansion.

shortest side of the image, and Δ is the offset value sampled from a half Gaussian distribution. The lightweight segmentation network is initialized for each image. The MobileNet v2 head is pre-trained on the ImageNet dataset [58]. We fix all the parameters of it and optimize the rest components for 50 steps using the ground truth generated by the scribble. We use the scribble annotations provided by [39]. In order to expand the labeled areas of these scribbles, as illustrated in Fig. 6, we over-segment the input image into superpixels by SLIC [59]. The class labels of these superpixels depend on the scribbles. For a superpixel SP_j, if there is a scribble SC_i with class c_i overlap SP_j, we mark all the pixels in SP_j with label c_i. The loss function we employed is the weighted cross-entropy loss:

$$CELoss_i = -\sum_{i=1}^{N} w_{c_i} y_{c_i} \log(p_i), \tag{5}$$

where N is the number of categories, p_i denotes the prediction probability of class c_i, and $y_{c_i} = 1$ indicates the ground truth of this prediction is c_i, while we calculate the weight for each category by:

$$w_{c_i} = \frac{n_{c_i}}{N_{all}}, \tag{6}$$

in which n_{c_i} is the number of the pixels of category c_i and N_{all} is the number of all the labeled pixels. We employ SGD [60] as our network optimizer with the learning rate set to 0.01. When the optimization is completed, the GDC samples extra 50 offset values to generate 50 different segmentation results. We average these 50 results as the final segmentation result.

2) Performance: a) The advantage of GDC: We first evaluate the performance of our GDC based lightweight network on the Pascal-Context dataset [9] which involves 59 categories of objects and stuff. The accuracy is evaluated by the overall accuracy and the mean Intersection-over-Union (mIoU) score. In order to demonstrate the advantage of the proposed GDC over the other convolutions, we use 3 other different convolution kernels, namely, normal convolution, dilated convolution [5] and deformable convolution [56], to replace the GDC layer in our lightweight network to produce the three alternative lightweight networks for comparison. Following the symbol conventions of the previous sections, the output of deformable convolution is calculated by

$$output^c = \sum_{e_i \in \mathcal{E}} w^{e_i} \cdot x^{c+(\Delta_{base}+\Delta_i)\odot e_i}, \tag{7}$$

where c is the coordinate of the center convolution position, w^{e_i} is the weight of the convolution kernel, and $x^{c+(\Delta_{base}+\Delta_i)\odot e_i}$ is the input feature vector at the coordinate $c + (\Delta_{base} + \Delta_i) \odot e_i$. The convolution position is mainly controlled by a bias Δ_i and this bias is obtained by a trainable neural network. When $\Delta_i = 0$, the deformable convolution degenerates into the dilated convolution.

As shown in the left part of Table I, we use a normal 3×3 convolution layer as the baseline. For the Pascal-Context dataset, the normal convolution baseline achieves 53.41% mIoU. The dilated convolution with factor $= 6$ and deformable convolution increase the mIoU to 60.01% and 54.67%, respectively, because they both can expand the receptive field. Notably, the deformable convolution can hardly boost the performance. The possible reason is that the low level feature maps extracted by the one image training based lightweight network is insufficient for learning a reliable offset. By contrast, our GDC achieves the highest mIoU of 65.12%, which is 11.71% higher than the baseline. We will discuss the reasons for this performance improvement later.

We also evaluate our method on the validation set of Pascal-VOC 2012 [10] which contains 1,449 images with 21 categories. We report the results in the right part of Table I. The normal convolution baseline attains the mIoU of 68.94%. The dilated convolution with dilated factor $= 6$ and deformable convolution achieve the mIoU values of 71.95% and 68.69%, respectively. Our GDC by comparison increases

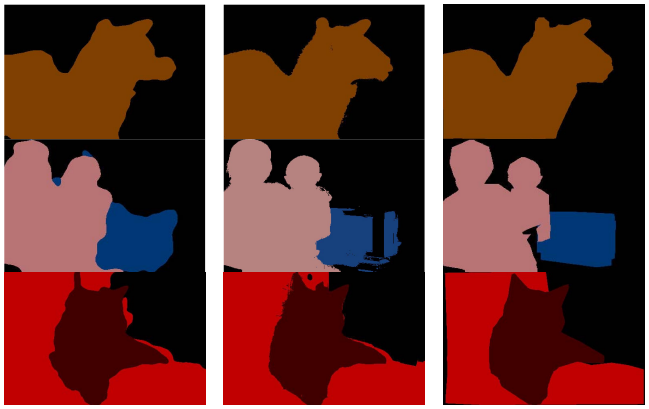


Fig. 7. Visualization of the ground truth on Pascal-VOC 2012.

TABLE II

PERFORMANCE COMPARISON OF THE SEGMENTATION MODELS TRAINED WITH THE PSEUDO LABEL ON PASCAL-VOC 2012 VALIDATION SET. † MEANS REFINING BY CRFS

Methods	**Conference**	**mIoU (%)**
ScribbleSup†	CVPR2016 [39]	63.1
RAWKS	CVPR2017 [40]	61.4
NormalizedCutLoss	CVPR2018 [61]	62.4
GraphNet-Initial	ACMMM2018 [49]	63.3
Ours	-	**64.9**

the mIoU to 74.06%, which is the highest on this dataset among the four lightweight segmentation networks.

b) The quality of the single image segmentation results: Following GraphNet [49], we use the augmented data by Hariharan *et al.* [62] to setup a weakly supervised segmentation experiment to evaluate the quality of our single image segmentation results. We use the single image segmentation results as the pseudo training label to train a DeepLab-v1 [20] on the training sets of Pascal-VOC 2012 and Pascal Context, respectively. The visualization of the pseudo ground truth and the real ground truth on Pascal-VOC 2012 training set is depicted in Fig. 7. We compare our method with the four state-of-the-art scribble supervised segmentation methods, i.e., ScribbleSup [39], RAWKS [61], NormalizedCutLoss [40] and GraphNet [49] on the validation subset. The performance on Pascal-VOC 2012 are reported in Table II. Our method achieves the best mIoU of 64.9%. Fig. 8 shows that the segmentation results generated by our weakly supervised model are comparable to those with strong supervision. This demonstrates the quality of the weakly labels produced by our GDC one image segmentation. Table III compares the mIoU results of our method with GraphNet-Initial [49] on Pascal Context validation set. It can be seen that our method also achieves better performance.

c) Ablation: We further investigate the effects of standard deviation Σ on the achievable performance of our GDC using Pascal-Context in Table IV. The value of Σ controls the overall scale of the receptive field. Comparing Table IV with

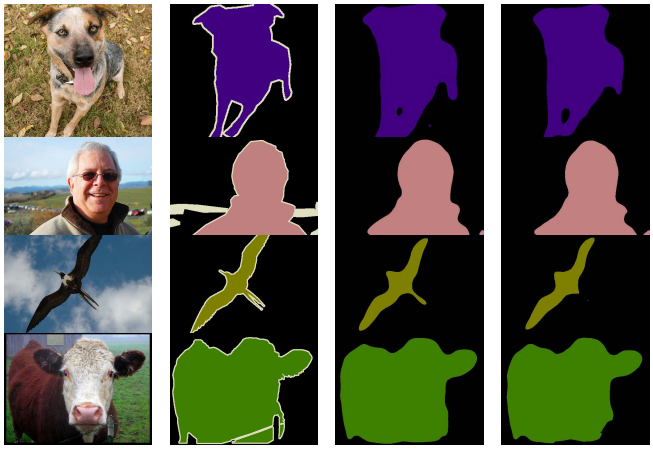


Fig. 8. Visualization results of ground truth, strong supervision and our weakly supervised segmentation on Pascal-VOC 2012 validation set.

TABLE III

PERFORMANCE COMPARISON OF TWO SEGMENTATION MODELS TRAINED WITH THE PSEUDO LABEL ON PASCAL-CONTEXT VALIDATION SET

Methods	**Conference**	**mIoU (%)**
GraphNet-Initial	ACMMM2018 [49]	33.1
Ours	-	**34.1**

TABLE IV

THE OVERALL ACCURACY AND MIOU WITH DIFFERENT Σ ON PASCAL CONTEXT DATASET

Σ	0.1	0.2	0.3
Overall Acc.	0.8591	**0.8646**	0.8602
mIoU (%)	63.97	**65.12**	64.88

Table I, it can be seen that our GDCs with different Σ values all achieve better performance than the three existing convolutions. In particular, with $\Sigma = 0.2$, the GDC achieves the best result. As Σ decreases or increases, there is a slight decrease in performance. The reason is that a too small Σ results in a small sampling area which is inefficient to aggregate large-scale information, while a too large Σ may fuse superfluous noise features.

We further discuss the influence of different feature extractors on single image segmentation. Similar to the operation of MobileNetV2, we select the output of 4x downsampling layer and 2x downsampling layer of different backbones as the result of feature extractor. As shown in Table V, MobileNetV2 achieves the highest ACC score of 0.9092, while VGG contributed the highest mIOU score of 80.34%. Experiments show that the complicated backbone is not a good choose for single image segmentation task.

B. Semantic Segmentation

1) Implementation: We implement the proposed GDPP module for semantic segmentation. First, its performance is

TABLE V

THE OVERALL ACCURACY AND MIOU WITH DIFFERENT BACKBONE ON PASCAL-VOC 2012 VALIDATION SET

BackBone	Overall Acc.	mIoU (%)
MobileNetV2	**0.9092**	74.06
Efficientnet-b0	0.5827	31.46
Resnet-50	0.9083	74.87
Resnet-101	0.9076	74.59
Xception	0.8439	69.46
DenseNet	0.8738	77.24
VGG	0.8797	**80.34**

TABLE VI

SEMANTIC SEGMENTATION RESULTS OF FOUR METHODS ON PASCAL-VOC 2012 VALIDATION SET

Methods	Setting	mIoU (%)	Params.
MobileNetv2		45.81	-
+ ASPP	50 epochs	63.01	33.6×10^5
+ ASPP	100 epochs	65.03 ▲2.02	33.6×10^5
+ S-ASPP	50 epochs	62.29	5.9×10^5
+ S-ASPP	100 epochs	63.16 ▲0.87	5.9×10^5
+ GDPP	50 epochs	62.44	5.9×10^5
+ GDPP	100 epochs	64.71 ▲2.27	5.9×10^5

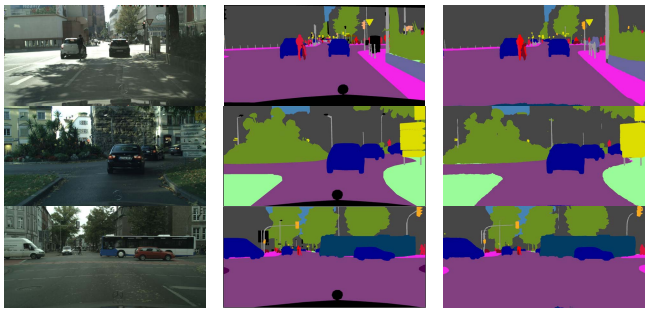


Fig. 9. Visualization results of our method on Cityscapes validation set.

TABLE VII

COMPARISON OF SEMANTIC SEGMENTATION PERFORMANCE ON CITYSCAPES VALIDATION SET

Methods	Backbone	mIoU (%)
PSPNet [6]	ResNet-101	77.8
PSANet [63]	ResNet-101	79.1
BiSeNet [64]	ResNet-101	80.3
DenseASPP [65]	DenseNet-121	76.6
DeeplabV3 [21]	ResNet-101	80.1
GDPP	ResNet-101	**80.6**

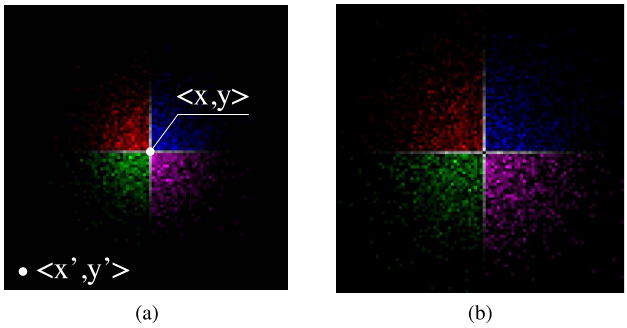


Fig. 10. The sample points of the GDC kernel at position $<x, y>$ with (a) $\Sigma = 0.1$, and (b) $\Sigma = 0.15$.

evaluated on Pascal-VOC 2012 dataset. The backbone network is a MobileNet v2 pre-trained on ImageNet dataset [58]. The feature maps generated by the last convolutional layer of the backbone network are sent to the GDPP module. We adopt one GDC layer and two dilated convolution layers in the GDPP module. The two dilated convolutions aggregate the small and very large scale information, respectively, while the GDC layer produces the rich middle-scale information. Following the setup in DeepLabs, the small and large dilated factors are set to 1 and 18, respectively. The base offset of our variant GDC is set to 9, and the Gaussian standard deviation Σ is chosen to be 2. Following the depthwise separable convolution in MobileNet, we modify all the convolution layers in the GDPP module to the separable mode to reduce the number of parameters. The tail segmentation network is the same as Deeplabv3 [21]. The optimizer is SGD with 0.028 learning rate which also follows the setup of the official DeepLabv3. The training epochs are set to 50 and 100, respectively. We use the original Pascal-VOC 2012 training dataset with 1464 images to accomplish the network optimization.

2) Performance: The semantic segmentation results of the four methods on Pascal-VOC 2012 validation set are reported in Table VI. We evaluate the MobileNet v2 backbone with the tail segmentation network and use the result as the baseline. This baseline network does not have any multi-scale information fusing module and can only get an mIoU value of 45.81%. First, we set the number of training epochs to 50. The original ASPP module [21] with 33.6×10^5 parameters can boost the mIoU to 63.01%. Our GDPP module achieves an mIoU of 62.44%, which is slightly lower than the ASPP, but it has only 5.9×10^5 parameters. To be fair, we also replace the GDC in the GDPP module by a separable dilated convolution and call the resulting module as S-ASPP. It can be seen from Table VI that this S-ASPP module gets an mIoU of 62.29% which is slightly lower than our GDPP module.

Next, we set the number of training epochs to 100 and train the three modules again. The results obtained are also presented in Table VI, where the corresponding increases in mIoUs over the results of 50-epochs training are also provided. It is worth noting that the advantage of the GDPP over the S-ASPP is clearly revealed. Specifically, the GDPP module achieves an mIoU of 64.71% which is 1.55% higher than the S-ASPP. We will discuss the reason in the discussion section.

3) More Dataset: We then evaluate our GDPP module on Cityscapes datasets. We use our GDC to replace the dilated convolution in the ASPP module of the original Deeplabv3. We keep the largest and smallest scales convolution layers as the same as the ASPP, and replace the two middle convolution

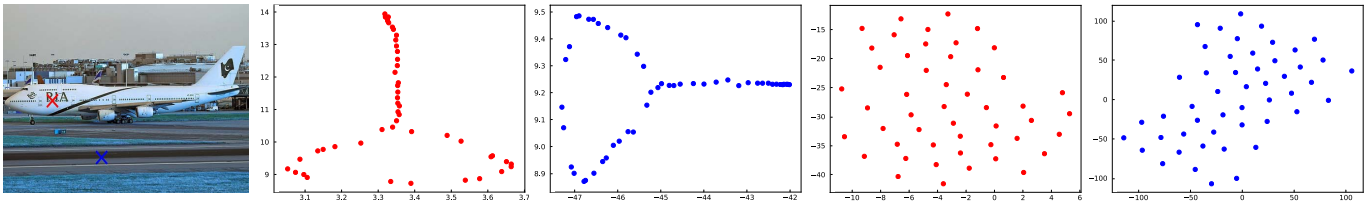


image (a) (b) (c) (d)

Fig. 11. We use the tSNE to embed the feature vectors at the red and blue crosses of the image. Red and blue points are the embedded feature vectors of the red and blue cross positions, respectively. (a) and (b) are produced by the dilated convolution, while (c) and (d) by our GDC.

layers by GDC. The basic offset of GDC is consistent with the dilated convolution. Here, the backbone network is Resnet101 [16] pre-trained on ImageNet dataset. The original DeeplabV3 achieves an mIoU of 80.1% on the validation set. Our GDPP module can boost the mIoU to 80.6%. The visualization results of our model can be found in Fig. 9. The comparison with more state-of-the-art models is shown in Table VII. It can be seen that our model achieves the highest mIoU.

VI. Discussion

As confirmed in Table I, our proposed GDC and the dilated convolution both outperform the normal convolution. Obviously, larger receptive field is beneficial for aggregating global contextual information. We now further discuss why the GDC outperforms the dilated convolution. Generally, the GDC can efficiently expand the receptive field without the need to stack many convolutional layers. The standard deviation Σ of the Gaussian distribution controls the overall scale of the receptive field extending. As an illustration, we initialize a GDC kernel and sample 100 times. In Fig. 10, we visualize the sample positions of the GDC kernel with two Σ values at the central position $<x, y>$. Clearly, a larger Σ leads to a larger receptive field.

The core idea of the GDC is to simulate the dynamic receptive field mechanism in the human being's visual system. Although the conventional dilated convolution can extend the receptive field, it can only use one or a group of fixed scales. This is totally different from the human visual system. By contrast, our GDC randomly selects different sampling positions to simulate this dynamic human receptive field. As illustrated in Table VI, such a dynamic convolution can achieve better performance than the conventional dilated convolution. The reason is that our dynamic convolution introduces more randomness into the training process. Different sample positions can fuse richer features in various scales. The tail segmentation network trained with these richer features can be optimized to a more robust status.

Moreover, vivid feature maps help to alleviate the problem of overfitting. The single image segmentation experiments of Subsection V-A offer the best evidence. Since there is only one image for optimizing the network, we need to prevent our lightweight segmentation network from overfitting into the scribble ground truth. Our dynamic convolution stochastically forms the convolution kernel of various spatial scales. In a way, it is equivalent to do some data augmentation in the feature space. For the sake of visual interpretation, we collect the feature vectors of the two positions, the red and blue crosses in the image of Fig. 11 in 50 optimizing steps, and use tSNE [66] to embed these feature vectors into two dimensions for visualizing. Visualizations of the embedded vectors are shown in the four scatter graphs in Fig. 11. Specifically, Figs. 11a and 11b depict the embedded feature vectors of the dilated convolution, while Figs. 11c and 11d show the feature vectors generated by our GDC. Obviously, the feature vectors produced by our GDC are more diverse than the dilated convolution. Thus, overfitting can be effectively alleviated.

Another discussion is why we sample the offset values from the Gaussian distribution. We believe that Gaussian distribution can simulate the correlation between feature vectors. For example, in Fig. 10a, the feature vector located at $<x', y'>$ has few relation with the feature vector located at $<x, y>$. If we aggregate the information at $<x', y'>$ into $<x, y>$, the feature map will become turbid. We design an simple experiment to illustrate this point. We replace the GDC in the lightweight segmentation network with a random dynamic convolution. The offset values of this random dynamic convolution are sampled from a uniform distribution. The result is shown in Table VIII. Observe that the random dynamic convolution achieves an mIoU of 68.10% which is even lower than the baseline.

Furthermore, we discuss the effect of GDC on stacking for the semantic segmentation task. We fix the largest and smallest scales convolution layers of GDPP and increase the number of middle scales layers n. As shown in Table IX, stacking the GDC layers can enhance the performance. However, it is

TABLE VIII
An Experiment on Pascal-VOC 2012 Validation Set

Template	Setting	mIoU (%)
Baseline	normal conv 3×3	68.94
GDC	$\Sigma = 0.2$	74.06
Random	-	68.10

TABLE IX
Comparative Experiment of GDC Stack

n	Δ_{base}	mIoU (%)
1	9	79.8
2	6 , 12	80.6
3	6 , 9 , 12	80.4

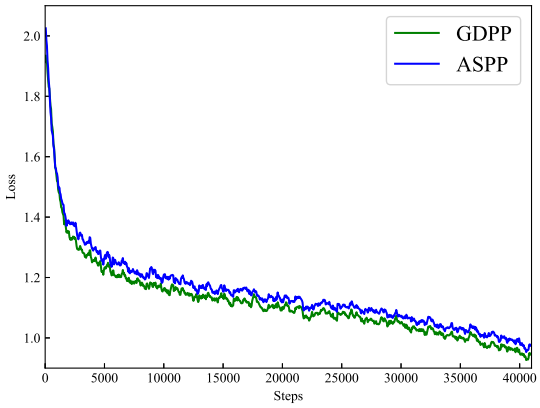


Fig. 12. Illustration of the training loss on Cityscapes training set.

not as more as better. It gets the promising performance when $n = 2$.

At last, we conduct experiments to show the stability of our convolution. We replace the ASPP module of DeeplabV3 with our GDPP, and make the stability comparison with the original DeeplabV3. We use two Gauss convolution layers in GDPP and remain other parameters unchanged. As shown in Fig. 12, the loss keeps declining steadily during our training with Gaussian convolution. It shows that our Gaussian convolution is stable enough.

VII. Conclusion

In this paper, we have propose a novel convolution GDC for the fast and effective single-image segmentation task. The proposed GDC can dynamically change its receptive field by sampling different offset values from a Gaussian distribution. This GDC can not only be employed for the single-image segmentation with scribbles but also be implemented for the common semantic segmentation network with a Gaussian Dynamic Pyramid Pooling. Our experiments have demonstrated that the GDC achieves better performance on the image segmentation than other existing forms of convolution, including dilated convolution and deformable convolution. In addition, the limitation of the proposed GDC is the difficulty to form a deep network. That means the GDC is more suitable for the fast single image segmentation than semantic segmentation. And we believe that the GDC can also help other computer vision tasks, such as fast classification and detection, which will be our future work.

References

[1] Y. Zhang, X. Sun, J. Dong, C. Chen, and Q. Lv, "GPNet: Gated pyramid network for semantic segmentation," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107940.

[2] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[3] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "GrabCut in one cut," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1769–1776.

[4] D.-J. Chen, H.-T. Chen, and L.-W. Chang, "SwipeCut: Interactive segmentation via seed grouping," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2959–2970, Sep. 2020.

[5] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 2881–2890.

[7] C. D. Gilbert and T. N. Wiesel, "Receptive field dynamics in adult primary visual cortex," *Nature*, vol. 356, no. 6365, p. 150, 1992.

[8] N. W. Daw and N. W. Daw, *Visual Development*, vol. 14. New York, NY, USA: Springer, 2006.

[9] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[11] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[12] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2014, pp. 1–14.

[15] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2016, pp. 770–778.

[17] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Feb. 5, 2021, doi: 10.1109/TCSVT.2021.3056725.

[18] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Feb. 2017.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[22] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[23] N. Gao, Y. Shan, Y. Wang, X. Zhao, and K. Huang, "SSAP: Single-shot instance segmentation with affinity pyramid," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 661–673, Feb. 2021.

[24] Q. Wang, C. Yuan, and Y. Liu, "Learning deep conditional neural network for image segmentation," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1839–1852, Jul. 2019.

[25] D. Lin, R. Zhang, Y. Ji, P. Li, and H. Huang, "SCN: Switchable context network for semantic segmentation of RGB-D images," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1120–1131, Mar. 2020.

[26] W. Ji, X. Li, F. Wu, Z. Pan, and Y. Zhuang, "Human-centric clothing segmentation via deformable semantic locality-preserving network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4837–4848, Dec. 2020.

[27] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.-T. Sun, "Cross-domain complementary learning using pose for multi-person part segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1066–1078, Mar. 2021.

[28] Z. Zhang *et al.*, "Tracking-assisted weakly supervised online visual object segmentation in unconstrained videos," in *Proc. 26th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 941–949.

[29] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, "Affinity attention graph neural network for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 25, 2021, doi: 10.1109/TPAMI.2021.3083269.

[30] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.
[31] X. Sun, C. Chen, J. Dong, D. Liu, and G. Hu, "Exploring ubiquitous relations for boosting classification and localization," *Knowl.-Based Syst.*, vol. 196, May 2020, Art. no. 105824.
[32] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4981–4990.
[33] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3791–3800.
[34] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2018, pp. 7024–7031.
[35] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1665–1674.
[36] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proc. Conf. Artif. Intell. (AAAI)*, 2020, pp. 12765–12772.
[37] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 549–565.
[38] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *The IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2018, pp. 616–625.
[39] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.
[40] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1818–1827.
[41] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1635–1643.
[42] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.
[43] S. Hong, S. Kwak, and B. Han, "Weakly supervised learning with deep convolutional neural networks for semantic segmentation: Understanding semantic layout of images with minimum human supervision," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 39–49, Nov. 2017.
[44] C. Redondo-Cabrera, M. Baptista-Rios, and R. J. Lopez-Sastre, "Learning to exploit the prior network knowledge for weakly supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3649–3661, Jul. 2019.
[45] C. Chen, X. Sun, Y. Hua, J. Dong, and H. Xv, "Learning deep relations to promote saliency detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 10510–10517.
[46] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4823–4836, Dec. 2020.
[47] X. Li, H. Ma, and X. Luo, "Weaklier supervised semantic segmentation with only one image level annotation per category," *IEEE Trans. Image Process.*, vol. 29, pp. 128–141, 2020.
[48] B. Xiong, S. D. Jain, and K. Grauman, "Pixel objectness: Learning to segment generic objects automatically in images and videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2677–2692, Nov. 2019.
[49] M. Pu, Y. Huang, Q. Guan, and Q. Zou, "GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation," in *Proc. ACM Multimedia*, 2018, pp. 483–491.
[50] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 697–707.
[51] B. Wang *et al.*, "Boundary perception guidance: A scribble-supervised semantic segmentation approach," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Aug. 2019, pp. 3663–3669.
[52] G. Wang *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.
[53] Z. Ji, Y. Shen, C. Ma, and M. Gao, *Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation* (Medical Image Computing and Computer Assisted Intervention). Cham, Switzerland: Springer, 2019, pp. 175–183.
[54] W. Lu, D. Gong, K. Fu, X. Sun, W. Diao, and L. Liu, "Boundarymix: Generating pseudo-training images for improving segmentation with scribble annotations," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107924.
[55] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12546–12555.
[56] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 764–773.
[57] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. Int. Conf. Learn. Represent.*, 2009, pp. 248–255.
[59] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[60] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Physica-Verlag HD, 2010, pp. 177–186.
[61] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2953–2961.
[62] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Apr. 2011, pp. 991–998.
[63] H. Zhao *et al.*, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
[64] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[65] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
[66] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Xin Sun (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the College of Computer Science and Technology, Jilin University, in 2007, 2010, and 2013, respectively. From 2016 to 2017, he did the post-doctoral research in computer science at the Ludwig Maximilian University of Munich. He currently works as a Researcher with the Technical University of Munich supported by the Humboldt Research Fellowship for Experienced Researchers. His current research interests include machine learning, computer vision, and remote sensing.

Changrui Chen was born in Shandong, China, in 1995. He received the bachelor's and master's degrees in computer science and technology from the Ocean University of China (OUC) in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree with the University of Warwick, U.K. His research interests are in computer vision and machine learning.

Xiaorui Wang was born in Henan, China, in 1995. He received the bachelor's degree in computer science and technology from the Ocean University of China (OUC) in 2019, where he is currently pursuing the master's degree with the Ocean Group of VisionLab. His main research is to solve computer vision problems with machine learning, especially interested in video object detection.

Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in applied mathematics from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, Edinburgh, U.K., in November 2003. He is currently a Professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision, and underwater image processing.

Huiyu Zhou received the Bachelor of Engineering degree in radio technology from the Huazhong University of Science and Technology of China in 1990, the Master of Science degree in biomedical engineering from the University of Dundee, U.K., in 2002, and the Doctor of Philosophy degree in computer vision from Heriot-Watt University, Edinburgh, U.K., in 2006. He is currently a Full Professor with the School of Informatics, University of Leicester, U.K.. He has published over 350 peer-reviewed articles in the field.

Sheng Chen (Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute in January 1982, the Ph.D. degree in control engineering from the City, University of London in September 1986, and the Doctor of Sciences (D.Sc.) degree from the University of Southampton in 2005. In September 1999, he joined the School of Electronics and Computer Science, University of Southampton, where he is currently a Professor in intelligent system and signal processing. He is a Chartered Engineer (C.Eng.) and a fellow of IET (FIET). He was elected to a fellow of the United Kingdom Royal Academy of Engineering in 2014.