Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Lightweight Semantic segmentation method based on GhostNet and Atrous Spatial Pyramid Pooling

To cite this article: Hengfeng Liao *et al* 2023 *J. Phys.: Conf. Ser.* **2477** 012080

View the article online for updates and enhancements.

You may also like

- Spine Image Segmentation Based on U-Net and Atrous spatial pyramid pooling
  Zhen Yang, Long Chen, Ting Fu et al.
- Research on Semantic Segmentation of Portraits Based on Improved Deeplabv3 +
  Kaihui Zhang, Xianhui Liu and Yufei Chen
- MMUU-Net: A Robust and Effective Network for Farmland Segmentation of Satellite Imagery
  Xumin Gao, Long Liu and Huaze Gong

ECS 244th Electrochemical Society Meeting

October 8 – 12, 2023 • Gothenburg, Sweden

50 symposia in electrochemistry & solid state science

▶ Deadline Extended!
Last chance to submit!

New deadline: April 21
submit your abstract!

This content was downloaded from IP address 119.13.223.52 on 19/04/2023 at 16:44

Lightweight Semantic segmentation method based on GhostNet and Atrous Spatial Pyramid Pooling

Hengfeng Liao*, Wei Yan, Deyu Liu

School of Computer and Information Sciences, Chongqing Normal University, Chongqing, China

2020110516005@stu.cqnu.edu.cn

Abstract: Semantic segmentation is frequently utilized in computer vision projects like remote sensing picture segmentation, unmanned driving, and medical image segmentation. A lightweight semantic segmentation model is suggested by taking into account three components of the network - parameters, calculation, and performance - to address the issue of deploying embedded platforms with constrained processing power and hardware storage. Dual Attention is used with Atrous Spatial Pyramid Pooling (ASPP) to obtain precise relevant data utilizing the lightweight network GhostNet as the foundation, and to lighten ASPP's computational burden, depthwise separable convolution is used. To obtain a distinct segmentation boundary, a multi-scale splicing technique is then used. With network parameters of 2.7810^{6}, a floating-point calculation of 1.931GFLOPs, and a MIoU of 72.13% in the experiments just on PASCAL VOC 2012, the model achieves an excellent compromise between computational efficiency and segmentation precision.

1. Introduction

FCN [1] opens the road of full convolution and end-to-end semantic segmentation. PSPNet [2] uses different pooling operations through space pyramid pooling to control the receptive field and improve the ability to obtain context information. DeepLab [3-5] integrates features of different expansion rates through serial cavity convolution, increasing the network's receptive field and successfully restoring the feature image's spatial information. BiSeNet [6], a traditional real-time semantic segmentation method, has produced effective results. In order to provide lightweight semantic segmentation, these models still need to be refined. The advantages of the lightweight convolutional neural network GhostNet [7] include its straightforward calculation, strong mobility, and lightweight structure. The advantages of lightweight convolutional neural networks include their easy calculations, excellent mobility, and lightweight construction.

Deep convolutional neural networks' performance is significantly enhanced by attention. Dual Attention originally appeared in DANet [8], where rich context dependencies on local characteristics were established using location attention and channel attention, which are akin to self-attention. This considerably improved the segmentation results. The network makes reasonable use of the attention, and performance improves with less calculation and parameter expansion.

On the basis of the aforementioned research, a lightweight semantic segmentation model is created that better considers the network's parameters, computation, and performance. The following are the main contributions:

- The limited networking In order to lessen the amount of parameters and calculations needed, GhostNet is employed as the network's backbone.

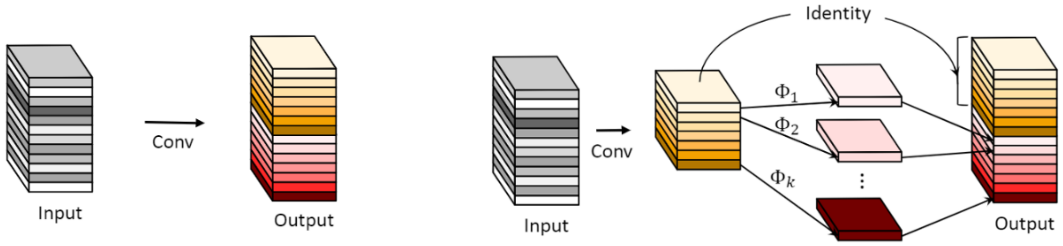
Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

- The lightweight backbone network frequently adversely affects the model's accuracy. ASPP and Dual Attention are employed to strengthen the processing of high semantic features to be able to capture context data of the model. The ASPP is simultaneously optimized in terms of minimizing the impact on the quantity of calculation and the number of parameters by switching from ordinary convolution to depthwise separable convolution.
- The method of multi-scale features splicing fully utilizes the valuable features produced by the backbone network, resulting in more accurate segmentation results for the model.

2. Correl~~ation theory~~

2.1. GhostNet

A unique Ghost module is proposed by GhostNet to produce more feature maps from inexpensive operations. Figure 1 depicts the Ghost module's central concept.



(a) The convolutional layer. (b) The Ghost module.

Figure 1. The Ghost module's fundamental concept

Examples of the standard convolution operation mode(a) and the modified ghost module(b). If we adopt the conventional way of thinking, we might believe that these identical feature maps contain duplicate data and that we should endeavor to prevent creating them. The ghost module only uses a portion of the input features to construct a significant number of new features in order to minimize computational cost when both approaches produce the same amount of feature maps. The bottleneck for Ghost is the stacking of Ghost modules, which makes it simple to create lightweight GhostNet.

2.2. ASPP

ASPP is a representative work of the DeepLab series, which is an effective way to capture context information. Figure 2 depicts its architecture.

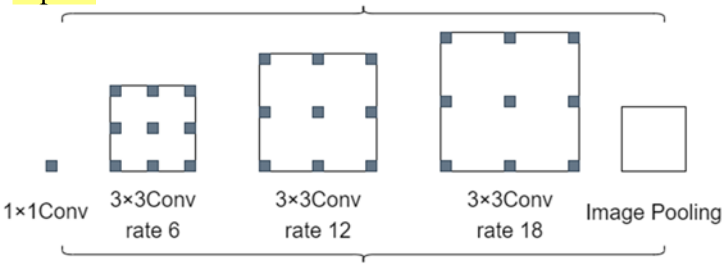


Figure 2. ASPP architecture

One 1×1 convolution kernel, three 3×3 convolution kernel with varying expansion rates to extract the image's multi-scale message, and one global average pooling layer make up the majority of the five branches that make up ASPP. In order to modify the amount of channels and upsampling, the output size of the global average pooling is 1×1, and the size of the input is restored by 1×1 convolution. The five branches are then joined together and spliced. When utilizing atrous convolution, the grid effect causes local information loss and a lack of correlation of long-distance information. ASPP solves these issues and can get feature information at various scales without the usage of pooling layers. As an alternative to conventional convolutions in ASPP, depthwise separable convolutions are employed in this study. L-ASPP is the name of the optimized module.

2.3. Dual Attention

Dual Attention is the combination of location attention and channel attention. Dual Attention can strengthen the features after passing the ASPP module so that its feature map pays more attention to some important information. Figure 3 shows the location attention and Figure 4 shows the channel attention.

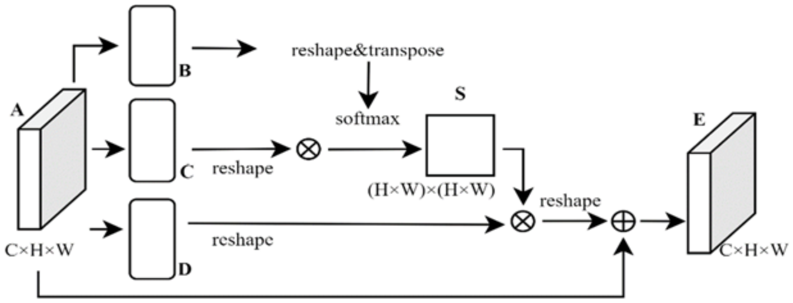


Figure 3. Location Attention

The location attention can be on location and capture the context information between any two locations.

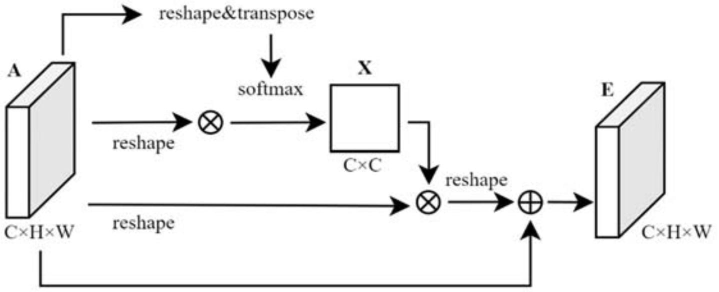


Figure 4. Channel Attention

The context information on the channel dimension can be captured by the channel attention. Each channel can be seen as a unique answer to a particular class in terms of high-level semantic properties. By enhancing the feature channels, the segmentation effect can be significantly enhanced.

3. Lightweight Semantic segmentation method based on GhostNet and ASPP

3.1. Overall framework

The backbone network selects the lightweight network GhostNet, and the high-semantic feature map of 32 times downsampling is processed through the optimized L-ASPP module and dual attention. The processed feature map of 32 times downsampling is then joined and fused with the feature maps of 16 times downsampling and 8 times downsampling, one after the other, using the multi-scale splicing method of the U-Net [9] architecture. Finally, it was shrunk to the extent of the segmentation prediction input image. In Figure 5, the overall framework is displayed.

The process of multi-scale splicing can maximize the reduction of detailed features while containing more rich spatial information. The model's reinforcement module at the base serves to increase the feature's global information, considerably enhancing the model's comprehension of segmentation accuracy. The framework of the model is straightforward and quick, and the focus of this paper is on lightweight and efficient semantic segmentation.

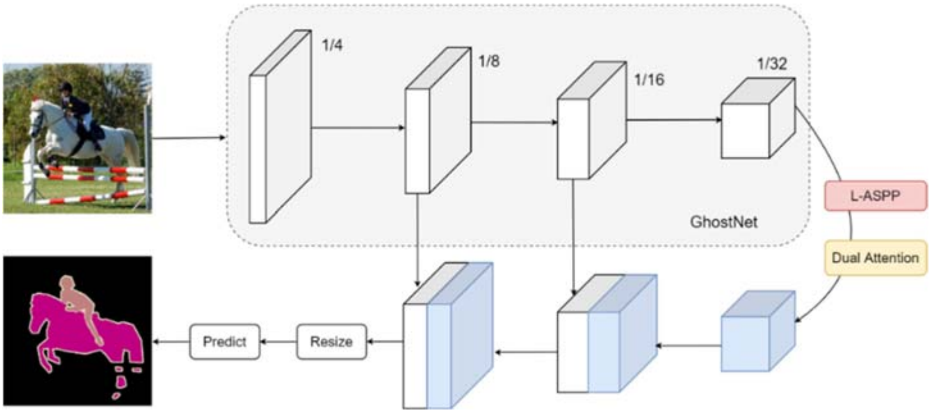


Figure 5. Overall framework

4. Experimental findings and evaluation

4.1. Experimental conditions

As the data set, the model is constructed and extensively studied using the extended PASCAL VOC 2012, which is frequently used in semantic segmentation. It consists of three sets of photos: a training set with an enlarged size of 10582 images, a validation set with a size of 1449 images, and a test set with a size of 1456 images. The system environment is Windows 11. The CPU is Intel i5-12400F, with 16 GB RAM. The GPU is a GeForce RTX3060 with 12 GB of video memory. Based on the open source PyTorch1.11.0, CUDA11.3.1, and CUDNN8.2.0.53 deep learning framework implementation.

4.2. Evaluation metrics

The amount of calculations (FLOPs) and the quantity of parameters describe the network's computational complexity (Params). The average ratio between the intersection and union of all categories is calculated to assess the effectiveness of the model. Mean Intersection over Union (MIoU) is typically employed as a measure of accuracy in semantic segmentation. The 21-category PASCAL VOC 2012 data set is used as an example, and the IoU for each category is computed independently. The category is denoted by n, the plus background class is denoted by (n+1), the official value is denoted by I the predicted value is signified by j, and the predicted I to j is signified by pij. The MIoU of a category can then be determined using the formula below:

$$MIoU=\frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{i=0}^{n}p_{ij}+\sum_{i=0}^{n}p_{ji}-p_{ii}} \qquad (1)$$

4.3. Data augmentation

During training, we set the crop size to 512×512, scaled the image, and distort the length and width. Gray bars were added to excess parts of the image. Flipped image; Gaussian blur.

Setting parameters: The skeleton GhostNet initialized the model parameters using pre-trained weights from ImageNet, which was able to shorten the model's training time and accelerate its convergence. The momentum technique (momentum=0.9) was paired with the SGD optimizer, and the learning rate was reduced by cos, following a cosine curve, with a high learning rate 0.007 and a low learning rate 0.00007. Additionally, the weight decay rate was 0.0001, which stopped the model from overfitting. The maximum output downsampling rate was 16, the batch size 8, and the training number 200 epochs in consideration of the problem of computational resources in the current context.

4.4. Experimentation Findings

For comparison, we chose some lightweight, enhanced versions of well-known real-time semantic segmentation models as well as some traditional semantic segmentation models. This study also discusses further optimization and enhancement, as well as some of the comparison models' more sophisticated concepts. Where the model's structure is represented by Model and Backbone. The outcomes from the experimental setting used in this paper include Params, GFOLPs, and MIoU.

On the PASCAL VOC 2012, Table 1 displays the experimental findings for the suggested model and other models.

Table 1. Results comparison using additional models on the PASCAL VOC 2012

Model	Backbone	Params	GFOLPs	MIoU/%
FCN-8s [1]	VGG16	134.68×10^6	321.343	62.48
BiSeNet [6]	ResNet18	49.43×10^6	13.651	65.34
（a）**PSPNet [2]**	MobileNetV2	2.89×10^6	5.873	68.47
（b）**PSPNet [2]**	ResNet50	46.72×10^6	118.128	79.02
（c）**DeeplabV3 + [5]**	MobileNetV2	5.82×10^6	52.846	71.72
（d）**DeeplabV3 + [5]**	Xception	54.71×10^6	166.342	76.97
（e）**HrnetV2_w18 [10]**		9.64×10^6	37.187	73.22
（f）**Ours**	GhostNet	2.78×10^6	1.931	72.13

According to Table 1, the suggested model performs better than the competition in terms of the amount of computations and the amount of parameters. After a lightweight backbone network is used in place of MobileNetV2, models (a) and (c) are contrasted. Our model shows clear advantages in terms of the quantity of parameters, the complexity of the calculations, and the accuracy, which also highlights the logic of the model design in this work. There are just 1.931 GFOLPs of floating point calculations. Additionally, since the calculation takes up less than 2 GFOLPs, embedded devices can benefit more from its use. The proposed model still has to be improved in terms of accuracy compared to models (b) and (d), which have good performance but significant overhead. The present model's 72.13% MIoU is likewise useful, especially since the amount of calculation required is less than 2% of (b) or (d). In summary, the suggested model enables lightweight and effective semantic segmentation while maintaining a good balance between the quantity of network parameters, computation, and performance.

4.5. Segmentation results

Three images from the test set are chosen at random for segmentation prediction. It is obvious that the suggested model can correctly identify the groups and satisfy the fundamental segmentation requirements. In comparison to other algorithms, our model segmented the initial image more precisely. People and bicycles are divided into comparable categories in the label image. Our model also segments the partially obscured horse leg in the second image. The third image detected the sofa class and effectively separated the cat and dog classes. The subdivision portion, however, still needs to be enhanced. Figure 6 displays the comparison's findings.

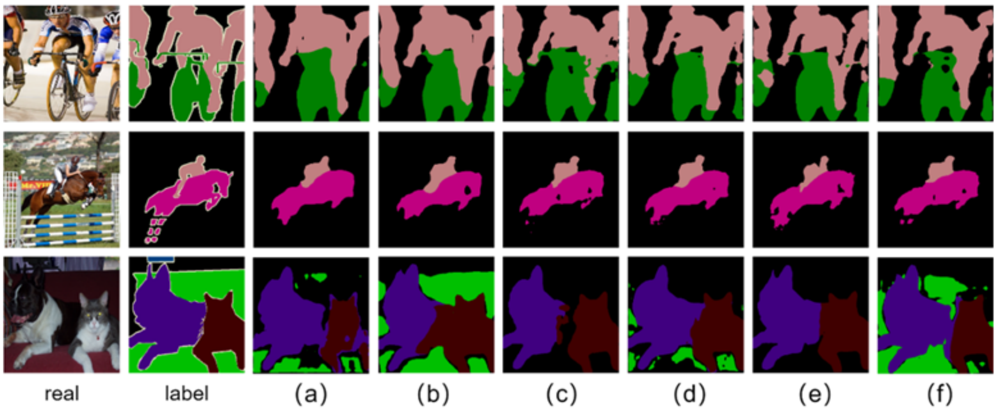


Figure 6. Results of several models' segmentation

5. Conclusion

The suggested model contains a few parameters and calculations thanks to the assistance of GhostNet as the supporting network and depthwise separable convolution. The L-ASPP module is optimized to capture a variety of context data with little overhead. The model can recover more precise information thanks to multi-scale concatenation. The experimental findings demonstrate that the suggested model has a good segmentation effect, great computational efficiency, and a smaller memory footprint. The demands of mobile and embedded devices are also met, and a fair balance between computing economy and segmentation accuracy is achieved.

References

[1] Long J., Shelhamer E., & Darrell T. 2015. Fully convolutional networks for semantic segmentation. *Proc. of the IEEE Conf. on computer vision and pattern recognition* pp. 3431-3440.

[2] Zhao H., Shi J., Qi X., Wang X., & Jia J. 2017. Pyramid scene parsing network. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*. pp. 2881-2890.

[3] Chen L. C., Papandreou G., Kokkinos I., Murphy K., & Yuille A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence,* 40(4), 834-848.

[4] Chen L. C., Papandreou G., Schroff F., & Adam H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587.*

[5] Chen L. C., Zhu Y., Papandreou G., Schroff F., & Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of the European Conf. on computer vision (ECCV)*. pp. 801-818.

[6] Yu C., Wang J., Peng C., Gao C., Yu G., & Sang N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *Proc. of the European Conf. on computer vision (ECCV).* pp. 325-341.

[7] Han K., Wang Y., Tian Q., Guo J., Xu C., & Xu C. 2020. Ghostnet: More features from cheap operations. *Proc. of the IEEE/CVF Conf. on computer vision and pattern recognition.* pp. 1580-1589.

[8] Fu J., Liu J., Tian H., Li Y., Bao Y., Fang Z., & Lu H. 2019. Dual attention network for scene segmentation. *Proc. of the IEEE/CVF Conf. on computer vision and pattern recognition.* pp. 3146-3154.

[9] Ronneberger O., Fischer P., & Brox T. 2015. U-net: Convolutional networks for biomedical image segmentation. *Int. Conf. on Medical image computing and computer-assisted intervention.* pp. 234-241.

[10] Sun K., Zhao Y., Jiang B., Cheng T., Xiao B., Liu D., Wang J., et al. 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514.*