AM-PSPNet: Pyramid Scene Parsing Network Based on Attentional Mechanism for Image Semantic Segmentation

Dikang Wu, Jiamei Zhao, and Zhifang Wang(✉)

Department of Electronic Engineering, Heilongjiang University, Harbin 150080, China
wangzhifang@hlju.edu.cn

Abstract. In this paper, AM-PSPNet is proposed for image semantic segmentation. AM-PSPNet embeds the efficient channel attention (ECA) module in the feature extraction stage of the convolutional network and makes the network pay more attention to the channels with obvious classification characteristics through end-to-end learning. To recognize the edges of objects and small objects more effectively, AM-PSPNet proposes a deep guidance fusion (DGF) module to generate global contextual attention maps to guide the expression of shallow information. The average crossover ratio of the proposed algorithm on the Pascal VOC 2012 dataset and Cityscapes dataset reaches 78.8% and 69.1%, respectively. Compared with the other four network models, the accuracy and average crossover ratio of AM-PSPNet are improved.

Keywords: Semantic segmentation · Efficient channel attention · Deep guide fusion

1 Introduction

In the field of computer vision, the application of neural networks mainly includes image recognition, target detection, and semantic segmentation. Semantic segmentation is the classification of each pixel in the image to determine the category of each point (such as background, person or car). Compared with image recognition and target location and detection, semantic segmentation not only provides the classification information of objects but also extracts the location information, which lays the foundation for other computer vision tasks [1, 2]. For example, in the field of autonomous driving, the system can automatically and quickly classify images to avoid obstacles. In medical image analysis, the semantic segmentation system automatically generates a simple disease report to help doctors diagnose. In precision agriculture, the machine performs semantic segmentation of crops and weeds in the image, realizes the weeding behavior of the machine, and accurately reduces the number of herbicides sprayed, greatly improving agricultural efficiency [3–5].

Traditional semantic segmentation methods generally divide images by extracting the grayscale information, texture shape, color and other shallow features of the image.

Y. Wang et al. (Eds.): ICPCSEE 2022, CCIS 1628, pp. 425–434, 2022.
https://doi.org/10.1007/978-981-19-5194-7_32

It divides the information of the same semantic category as the same region by fixing a range that has the same semantic category. The traditional methods are threshold segmentation, edge detection and region segmentation [6]. When the background is complex and contains multiple objects, the segmentation effect of traditional methods is not obvious, and the segmentation result is rough.

With the improvement of computer performance and the rapid development of deep learning in the field of computer vision, many image semantic segmentation methods based on deep learning have been proposed [7, 8]. The output of the full convolutional network (FCN) [9] is changed from a two-dimensional vector of fixed length to a two-dimensional space feature graph. The FCN adds an upsampling structure and then predicts each pixel. However, there are problems such as pixel loss in the upsampling process, resulting in rough segmentation results. To solve this problem, SegNet [10] was proposed and restored image details by retaining the maximum index during decoding. U-Net [11], the low-resolution features obtained by the encoder and the high-resolution features extracted by the decoder end are fused at the upsampling stage of the feature map to restore the refined features of the object and refine the edge information of the object. However, the above networks do not pay more attention to spatial context information, resulting in complex semantic information in the image, which easily confuses the target. To solve this problem, DeepLabv3 [12] uses atrous convolution of different expansion rates, and PSPNet [13] uses multiscale pooling features to aggregate multiscale contextual information. However, atrous convolution may lose some pixel position information, and PSPNet may cause information loss, making segmentation inaccurate. Therefore, the proposed DeepLabV3 + [12] in 2008 improved DeepLabV3 by introducing low-level features in the decoding stage. Nevertheless, the use of shallow information is very important in the process of feature fusion, and the segmentation effect is not obvious due to the insufficient use of shallow information. Excessive use of shallow features may lead to information redundancy. Semantic segmentation algorithms that treat all pixels equally are obviously different from human visual mechanisms. To enhance the influence on the region of interest in images and reduce information redundancy, researchers use an attention mechanism as the main method to solve such problems [14–17].

In this paper, a network model based on the attention mechanism AM-PSPNet is proposed. In this model, PSPNet is the backbone network, and the ECA attention module is added in the encoding stage, which can effectively learn the channel attention of each convolution block, reduce the noise and weight the feature channels, and improve the feature extraction performance of the network. In the decoding stage, the DGF module uses deep features to guide the expression of shallow features, strengthen the learning of important features, and restore the shallow features of image edge and texture information to achieve better pixel location and finer details.

2 AM-PSPNet

Based on PSPNet, AM-PSPNet is proposed in this paper. The ECA module and DGF module are added into the model in the encoding stage and decoding stage, respectively, which improves the feature extraction ability of the network and refines the classification results. The structure of the entire network is shown in Fig. 1.

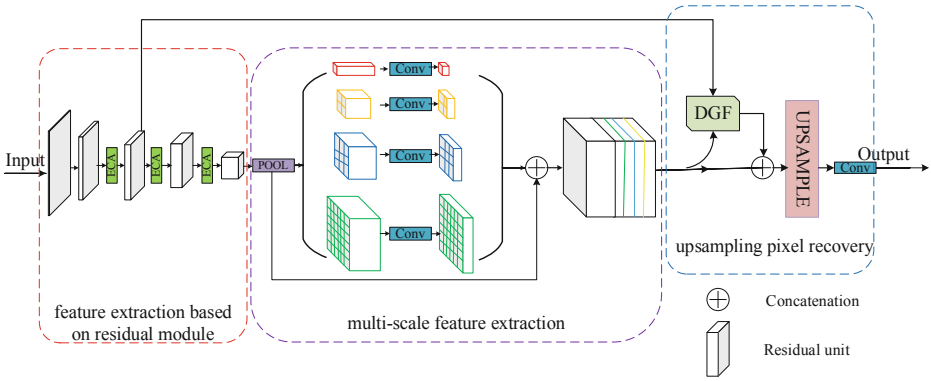

Fig. 1. AM-PSPNet framework.

AM-PSPNet is composed of three subnetworks: feature extraction based on a residual module, multiscale feature extraction and upsampling pixel recovery. The feature extraction subnetwork uses ResNet50 as the basic feature extraction. The network has five convolution modules of different structures. To avoid damage to the original ResNet structure, this paper adds ECA attention after the third, fourth and fifth convolution modules of ResNet so that the network can extract the discriminant features of images in the channel dimension. The multiscale feature extraction subnetwork uses a pyramid pooling module (PPM) to aggregate the context information of the multiscale to obtain the global context information. The DGF module is used in the upsampling recovered pixel subnetwork to guide the classification of shallow features more accurately through global contextual information.

2.1 Efficient Channel Attention Module

Adding the attention module to the existing convolutional neural network can bring performance improvement [18]. Most existing methods focus on more complex attention modules for better performance but result in increased computational burden on the network. To balance the relationship between network performance and complexity, the ECA module is introduced in this paper. The ECA module [19] adds little algorithm complexity while increasing network performance.

The ECA module has an improvement on the squeeze-and-excitation (SE) module [20], and the SE module can learn the channel attention of each convolutional block, which brings significant performance improvement to the deep convolutional neural network architecture. The SE module is used to control the complexity of the network, but dimension reduction can have a negative effect on predicting channel attention, and it is not necessary to obtain dependencies between all channels [21]. As a result, the ECA module efficiently captures local cross-channel interactions. As shown in Fig. 2, the ECA module implements global average pooling between channels without dimension reduction. It captures local cross-channel interactions through fast one-dimensional convolution of kernel size k. k is the coverage of the cross-channel interaction. Then, the sigmoid function is used to generate the weight proportion of each channel. The channel

attention feature is obtained by multiplying the given input by the channel weight. The size k can be determined by the adaptive function according to the size of the input channel C, and its calculation equation can be expressed as

$$k = \Phi(C) = \left| \frac{\log_2 c}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

$$C = \varphi(k) = 2^{(\gamma * k - b)} \tag{2}$$

In the equation, $|t|_{odd}$ is the odd number closest to t, the constant r is set to 2, and the constant b is set to 1.

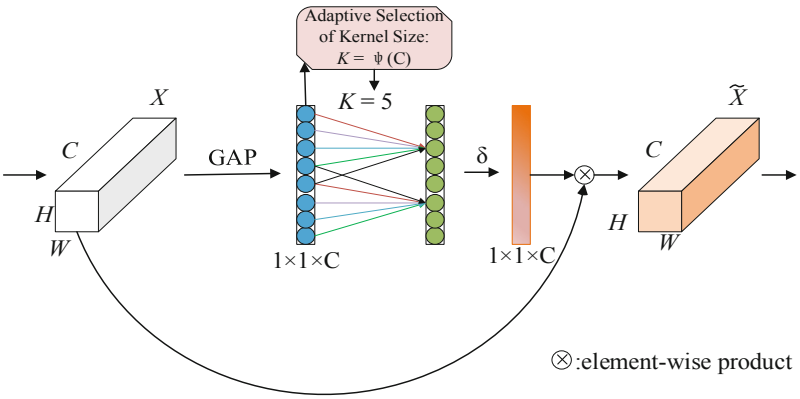

Fig. 2. ECA module.

2.2 Deep Guidance Fusion Module

Usually, multiscale context information is extracted by PPM. However, multistage spatial pooling will lose many fine information. Therefore, features in the deep layer of the network have strong semantic expression ability but poor pixel accuracy, while shallow features contain more pixel information. The direct superposition of deep features and shallow features easily produces considerable noise, while the segmentation accuracy of the model is reduced.

This paper proposes the deep guidance fusion module. As shown in Fig. 3, the DGF is embedded behind the PPM, and it performs global average pooling on deep features to produce attention maps. The shallow features are convolved with 3×3 to reduce the feature mapping channels from the CNN. Then, the shallow features are multiplied by the global attention force to screen out effective information. Finally, the output is added to the deep feature elements and upsampling to produce the final prediction results. To reconcile the contradiction between improving performance and reducing complexity, the output of the third stage is selected as a shallow feature in the feature extraction stage after many experiments.

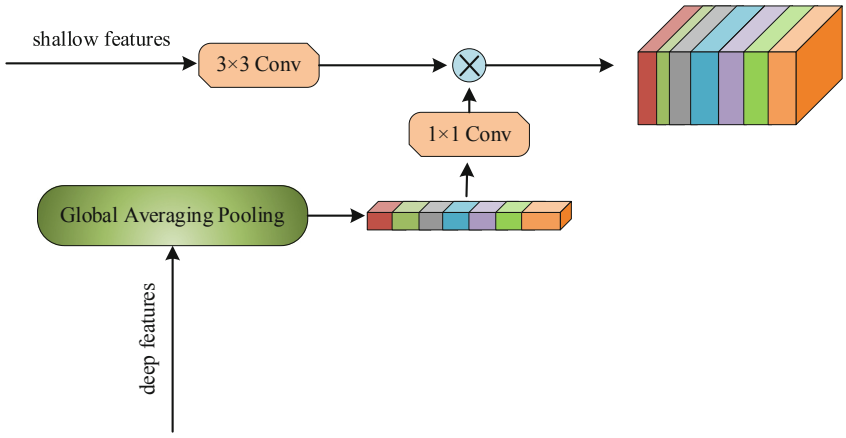

Fig. 3. DGF module.

3 Experiments and Analysis

3.1 Experimental Design

The PASCAL VOC 2012 dataset and the Cityscapes dataset are used to evaluate the performance of AM-PSPNet. First, ablation studies are carried out on the PASCAL VOC 2012 dataset, and then experiments are carried out on two datasets to compare the performance of the network.

To better reflect the performance of the model. Pixel accuracy (PA) is the ratio of correctly segmented pixel points to total pixel points, and mean intersection over union (mIoU) is the ratio of intersection and union of ground truth and prediction graph. The calculation equations are as follows:

$$PA = \frac{\sum_{i=o}^{N} n_{ii}}{\sum_{i=o}^{N} \sum_{j=0}^{N} n_{ij}} \tag{3}$$

$$mIoU = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{n_{ii}}{\sum_{j=1}^{N} n_{ij} + \sum_{j=1}^{N} (n_{ji} - n_{ii})}\right) \tag{4}$$

where N is the number of category labels and n_{ji} is the total number of pixels of true category i but predicted category j. n_{ii} and n_{ij} are similar to n_{ji}.

3.2 Ablation Study

To test the importance and performance of each part of the model, an ablation study is designed. To simplify ablation studies, all methods are performed on the PAS CAL VOC 2012 dataset using the same experimental environment to compare performance in different configurations. Resnet-50 is used as the feature extraction network in this paper. The clipping size of the input data is set to 380 × 380, and the batch size is set

Table 1. Ablation study on the PASCAL VOC 2012 dataset

Model	Stage2	Stage3	Stage4	mIoU(%)
ResNet-50				77.4
ResNet-50 + DGF				77.9
ResNet-50 + DGF	√			78.1
ResNet-50 + DGF		√		78.4
ResNet-50 + DGF			√	78.2
ResNet-50 + DGF		√	√	78.5
ResNet-50 + DGF	√	√	√	78.8

to 8. The performance of each module is compared in a fair way, and the corresponding results are shown in Table 1.

In this paper, feature extraction is divided into five stages. Stages 2, 3 and 4 in Table 1 indicate whether to add the ECA attention module in the second, third and fourth stages of network feature extraction, respectively, and " + DGF" indicates that the DGF module is added in the network decoding stage. Table 1 shows that the addition of the DGF module is beneficial to the improvement of network performance. When the ECA attention module is added to the second, third and fourth stages of network feature extraction, the network performance is the best.

3.3 Performance Evaluation on PASCAL VOC 2012

The validity of AM-PSPNet is verified using PASCAL VOC 2012, which is a public standard dataset commonly used in the field of semantic segmentation. It contains 1464, 1456 and 1449 images used for training, testing and verification, respectively. There are four categories of human, animal, vehicle and indoor objects, and there are 20 categories in total. There are 21 semantic categories, including one background category.

To accurately measure the performance of the model, AM-PSPNet, FCN-8S, U-Net, PSPNet and DeepLabV3 are experimentally verified on the PASCAL VOC 2012 dataset. The prediction results are shown in Table 2 and Table 3.

Table 2. Semantic segmentation results on the PASCAL VOC 2012 dataset

Model	PA (%)	mIoU (%)
FCN-8s	90.5	64.6
U-Net	91.8	70.4
DeepLabV3	94.3	77.8
PSPNet	94.2	77.4
AM-PSPNet	94.6	78.8

Table 3. Each category results on the PASCAL VOC 2012 testing set.

Model	FCN-8s	U-Net	DeepLabV3	PSPNet	AM-PSPNet
Background	90.3	90.7	93.4	93.3	93.8
Aeroplane	79.4	81.2	88.5	89.8	92.9
Bicycle	35.2	37.8	44.2	42.3	43.3
Bird	74.2	83.9	90.5	91.6	89
Boat	61.2	62.2	70.2	73.2	74.6
Bottle	61.4	68.2	80.6	77.9	79.6
Bus	79	91	91.8	90.1	92.3
Car	77.2	80.2	88.3	86.6	90.8
Cat	79.6	83.6	93.2	91.6	94
Chair	27.5	32.8	43.8	42.6	38.2
Cow	65.9	79.4	88.1	87.2	89.4
Diningtable	47	56.5	55.6	56.6	56.9
Dog	71.6	80.3	89.2	86.8	89.8
Horse	63.7	74.9	88.1	84.1	87.9
Motorbike	74.2	79.5	85.7	85.6	87.4
Person	79.2	81.7	86.3	86	86.9
Pottedplant	48.3	56.7	65.2	66.8	66.9
Sheep	69.3	78.6	85.2	85.3	88.4
Sofa	38.5	37.9	50.6	51.4	51.1
Train	72.2	85.4	82.8	85.9	85.1
Tvmonitor	62.6	58.3	73.9	72.9	76.7
mIoU	64.6	70.5	77.8	77.4	78.8

As seen from Tables 2 and 3, PSPNet achieves good prediction results compared with other semantic segmentation models, but AM-PSPNet achieves better prediction results; the PA is 94.6%, and the mIoU is 78.8%, which are 0.4% and 1.4% higher than the prediction results of PSPNet. AM-PSPNet obtains the highest accuracy for 15 of all categories of segmentation results. Compared with PSPNet, the segmentation results of 19 categories are improved, among which the segmentation effect of object categories with indistinguishable boundaries is significantly improved. For example, the segmentation results of the network for horse and sheep categories improved by 3.8% and 3.1%, respectively, compared with PSPNet. The use of the ECA module enhances the feature class resolution of the network, and the DGF module is helpful in restoring the image edge detail features. The experiment verifies the effectiveness of these two modules in AM-PSPNet.

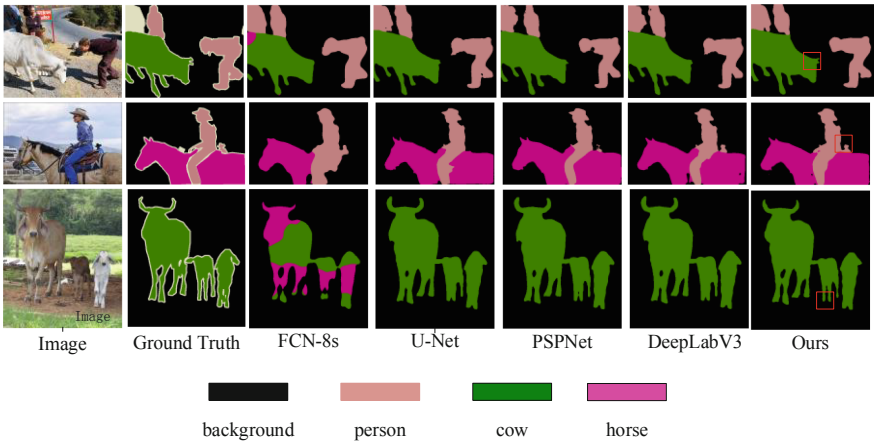

Fig. 4. Comparison of prediction results.

To display the segmentation effect of the model more intuitively, the comparison of the visualization results of each model is shown in Fig. 4. By observing the first line, it can be seen that PSPNet segmentation of cow horns is rough, while AM-PSPNet better retains segmentation details and makes the prediction more accurate and clearer. Compared with the picture in the second line, PSPNet failed to predict the distant figure completely, and several network models missed segmentation with serious loss of details. AM-PSPNet accurately expresses the details of the image. Compared with the picture in the third line, it can be seen that AM-PSPNet has a more delicate prediction of cattle legs. Compared with FCN-8s, U-Net, DeepLabV3 and PSPNet, the overall predicted contour of AM-PSPNet is smooth and delicate, and the predicted result is closer to the ground truth.

3.4 Performance Evaluation on Cityscapes

This paper also evaluated AM-PSPNet on the Cityscapes Dataset, which has 5000 images of driving scenes in urban environments, with 19 categories, recording street scenes in 50 different cities.

Resnet-50 is used as the backbone feature extraction network for network training. Affected by the GPU memory capacity, 380 × 380 is selected as the cutting size of the input in this paper, and the batch size is set to 6. The prediction results are shown in Table 4. The Am-pspnet proposed in this paper is superior to other networks, with mIoU reaching 69.1% and PA reaching 95.2%, improving by 1.6% and 1.1%, respectively, compared with the original network.

Table 4. Semantic segmentation results on the Cityscapes dataset

Model	PA (%)	mIoU (%)
FCN-8s	90.6	55.6
U-Net	92.4	61.7
DeepLabV3	94.7	68.7
PSPNet	94.1	67.5
AM-PSPNet	95.2	69.1

4 Conclusions

This paper uses AM-PSPNet as the backbone network, and the DGF module is proposed to guide shallow feature expression through deep features and achieve better pixel positioning. The ECA module is added in the feature extraction stage to improve the performance of the convolutional neural network architecture by learning the channel attention of each convolutional block. The experiment is carried out on the PASCAL VOC 2012 dataset and Cityscapes dataset. That, AM-PSPNet has good performance compared with FCN-8s, U-Net, PSPNet and DeepLabV3.

References

1. Wang, J., Liu, B., Xu, K.: Semantic segmentation of high-resolution images. Sci. China Inf. Sci. **60**(12), 1–6 (2017). https://doi.org/10.1007/s11432-017-9252-5
2. Yan, B., Niu, X., Bare, B., Tan, W.: Semantic segmentation guided pixel fusion for image retargeting. IEEE Trans. Multimedia **22**, 676–687 (2020)
3. Zhao, Y., Qi, M., Li, X., Meng, Y., Yu, Y., Dong, Y.: P-LPN: toward real time pedestrian location perception in complex driving scenes. IEEE Access **8**, 54730–54740 (2020)
4. Cheng, Z., Qu, A., He, X.: Contour-aware semantic segmentation network with spatial attention mechanism for medical image. Vis. Comput. **38**(3), 749–762 (2021). https://doi.org/10.1007/s00371-021-02075-9
5. Zhang, R., Chen, J., Feng, L., Li, S., Yang, W., Guo, D.: A refined pyramid scene parsing network for polarimetric SAR image semantic segmentation in agricultural areas. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)
6. Bai, S., Wang, C.: Information aggregation and fusion in deep neural networks for object interaction exploration for semantic segmentation. Knowl. Based Syst. **218**, 106843 (2021)
7. Hao, S., Zhou, Y., Zhang, Y., Guo, Y.: Contextual attention refinement network for real-time semantic segmentation. IEEE Access **8**, 55230–55240 (2020)
8. Ji, J., Lu, X., Luo, M., Yin, M., Miao, Q., Liu, X.: Parallel fully convolutional network for semantic segmentation. IEEE Access **9**, 673–682 (2021)
9. Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 640–651 (2015)
10. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2481–2495 (2017)

11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
13. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Computer Society (2016)
14. Lin, Z.K., Sun, W., Tang, B., Li, J.D., Yao, X.Y., Li, Y.: Semantic segmentation network with multipath structure, attention reweighting and multiscale encoding. Vis. Comput. 1−12 (2022). https://doi.org/10.1007/s00371-021-02360-7
15. Li, H., Qiu, K., Chen, L., et al.: SCAttNet: semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. IEEE Geosci. Remote. Sens. Lett **18**(5), 905–909 (2021)
16. Xia, Z., Kim, J.: Mixed spatial pyramid pooling for semantic segmentation. Appl. Soft Comput. **91**, 106209 (2020)
17. Wang, Z., Wang, J., Yang, K., Wang, L., Su, F., Chen, X.: Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. Comput. Geosci. **158**, 104969 (2022)
18. Yin, J., Xia, P., He, J.: Online hard region mining for semantic segmentation. Neural Process. Lett. **50**(3), 2665–2679 (2019). https://doi.org/10.1007/s11063-019-10047-3
19. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11531–11539 (2020)
20. Jie, H., Li, S., Gang, S., Albanie, S.: Squeeze-and-excitation networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
21. Wang, Y.-N., Tian, X., Zhong, G.: FFNet: feature fusion network for few-shot semantic segmentation. Cogn. Comput. **14**(2), 1–12 (2022). https://doi.org/10.1007/s12559-021-09990-y