Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8861886, 9 pages
https://doi.org/10.1155/2020/8861886

Research Article

An Encoder-Decoder Network Based FCN Architecture for Semantic Segmentation

Yongfeng Xing,[1,2] **Luo Zhong**,[1] **and Xian Zhong**[1]

[1]*School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China*
[2]*School of Software, Nanyang Institute of Technology, Nanyang 473000, China*

Correspondence should be addressed to Yongfeng Xing; xingyongfeng@163.com

Received 19 April 2020; Revised 29 May 2020; Accepted 9 June 2020; Published 7 July 2020

Academic Editor: Wei Wang

In recent years, the convolutional neural network (CNN) has made remarkable achievements in semantic segmentation. The method of semantic segmentation has a desirable application prospect. Nowadays, the methods mostly use an encoder-decoder architecture as a way of generating pixel by pixel segmentation prediction. The encoder is for extracting feature maps and decoder for recovering feature map resolution. An improved semantic segmentation method on the basis of the encoder-decoder architecture is proposed. We can get better segmentation accuracy on several hard classes and reduce the computational complexity significantly. This is possible by modifying the backbone and some refining techniques. Finally, after some processing, the framework has achieved good performance in many datasets. In comparison with the traditional architecture, our architecture does not need additional decoding layer and further reuses the encoder weight, thus reducing the complete quantity of parameters needed for processing. In this paper, a modified focal loss function is also put forward, as a replacement for the cross-entropy function to achieve a better treatment of the imbalance problem of the training data. In addition, more context information is added to the decode module as a way of improving the segmentation results. Experiments prove that the presented method can get better segmentation results. As an integral part of a smart city, multimedia information plays an important role. Semantic segmentation is an important basic technology for building a smart city.

1. Introduction

Convolution neural network is the part and parcel of image recognition, detection, and segmentation. The image semantic segmentation can provide a strong foundation for the construction of a smart city and has received much attention and research in recent years. Semantic segmentation is aimed at classifying all pixels in the image according to a specific category, which is commonly referred to as dense prediction. It is different from image classification because we do not classify the entire image into one class but all pixels. Thus, we boast a set of predefined categories and we need to distribute a tag to all pixels of the image according to the context of various objects in the image [1]. Deep neural network is no secret to the innovation of computer vision, particularly image classification. Since 2012, it has surpassed its predecessors by a large margin. In fact, artificial intelligence is superior to human in image classification. Inevitably, we adopted the same technology for semantic segmentation. Therefore, we put forward a network structure on the basis of encoder-decoder and atrous spatial pyramid pooling [2]. At the same time, a combination of multiple loss functions is used to be the ultimate loss function.

A relatively naive approach to construct the neural network architecture is simply stacking several convolutions, using the same padding to preserve that the dimensions remain the same and then output an ultimate segmentation map. Through a series of feature mapping transformations, the corresponding mapping of segmentation results can be learned directly from the input image. But it is quite expensive in computation to keep the whole resolution in the whole network. This architecture is illustrated in Figure 1.

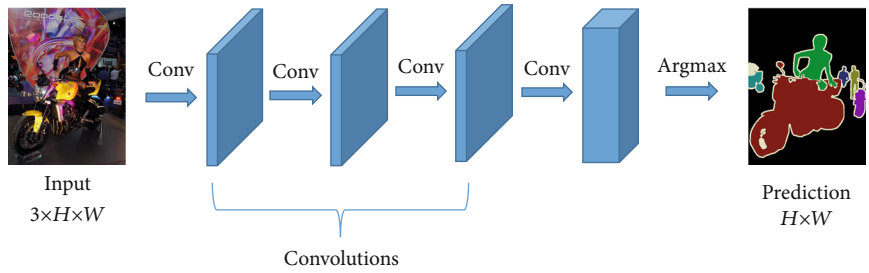


Figure 1: A simple method of constructing a neural network structure.

2. Related Works

In the deep convolution networks, the first layer studies the low-level notions, and the second layer studies the high-level feature mapping. As a way of maintaining the expression ability, the quantity of feature maps (channels) is usually increased while deepening the network. Different from the image classification which only needs the target category, image segmentation needs the location information of each pixel, so it cannot use pooling or trided convolutions to reduce the computation as safely as in the classification task. Image segmentation needs a whole-resolution semantic prediction. A popular image segmentation model is based on an encoder-decoder structure. In the encoder part, down sampling is adopted to reduce the input spatial resolution, so as to generate a lower resolution feature mappings (which is computationally efficient and can effectively distinguish different categories); in the decoder part, these feature representations are up sampled and restored to the full-resolution segmentation map.

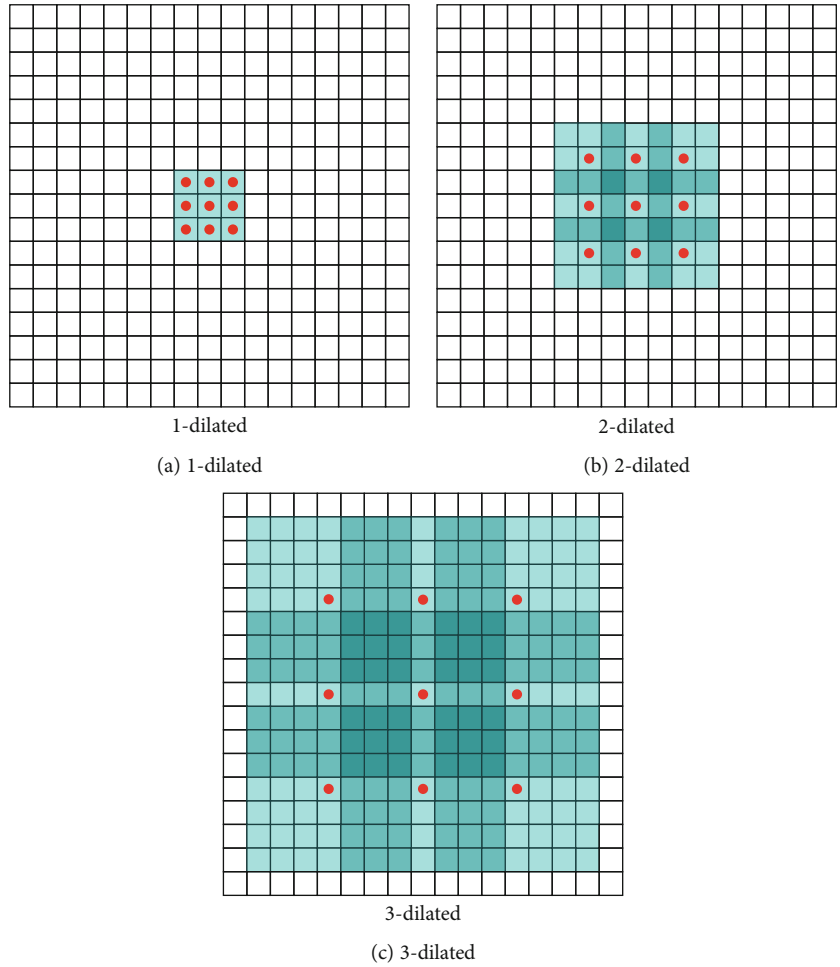
2.1. Fully Convolutional Network. Long et al. introduces the way to utilize end-to-end, pixel-to-pixel image segmentation task trained by the fully convolutional network at the end of 2014. In this paper, the author proposes to use the existing and well-researched image classification network as the encoder module of the network, adds transpose convolution layer in the decoding module, and upgrades the coarse feature mapping to the full-resolution segmentation mapping [3]. Full convolution network (FCNs) has achieved great success in the application of dense pixel prediction in semantic segmentation. The algorithm is required for predicting a variable for all pixels of the input image, a basic task in advanced computer vision understanding [1, 3]. Some of the most attractive applications include automatic driving [4], human-computer interaction [2, 5, 6], intelligent transportation system [7], auxiliary photo processing [8], and medical imaging [9]. The great achievements of FCNs come from the powerful characteristics picked up by CNNs. It is important that the convolution computer system makes the calculation efficiency of training and reasoning very high.

2.2. Encoder-Decoder. The encoder-decoder structure is a common architecture of current semantic segmentation algorithms. The structure is composed of an encoder and decoder. Classic image semantic segmentation algorithms such as FCN, U-net, and DeepLab all adopt this structure. The encoder is usually a network (VGG, Resnet, Xcepiton, etc); it consists of a deconvolution layer and upper sampling layer. Down sampling is aimed at capturing semantic or context information, while up sampling is aimed at recovering spatial information. Common decoders include bilinear interpolation, deconvolution, and dense up sampling convolution.

2.3. Dilated Convolution. In FCNs, because of continuous max pooling and down sampling operations, the feature resolution is greatly reduced. Finally, the feature mapping recovered by up sampling loses the detail sensitivity of the input image. In the full convolution network, the extended convolution is used instead of the standard convolution, so that the convolution network can accurately control the resolution of the image when calculating the feature response [10]. At the same time, the receiving the field of the filter is effectively expanded without adding the quantity of parameters and the amount of computing. Many experiments show that the algorithm uses more context information to obtain more dense features, thus improving the image semantic segmentation accuracy. It can be seen from Figure 2 that this is an expansion convolution filter with three different expansion rates: each element in the filter is a (a) 1-expansion convolution and a 3×3 receptive field, (b) 2-expansion convolution and a 7×7 receptive field, and (c) 3-expansion convolution and a 15×15 receptive field. The quantity of parameters related to each layer is the same. The receptive field increases exponentially and the number of parameters increases linearly [11].

Under the same size of convolution kernel, the receiving field of the convolution kernel can be increased by increasing the input stripe, as shown in Figure 3.

FCNs is a kind of deep convolution neural network, which has achieved good performance in pixel-level recognition tasks, but it still faces challenges in this changing and complex world. FCN is not a fully connected layer. The original method is to use the same size convolution layer stack as a way of mapping the input image to the output image. It produced strong results, but it was very expensive, because they cannot utilize any subsampling or pooling layers, because this will screw up the location of the instance. As a way of maintaining the resolution of the image, they must add many layers in a way that learns the low-level and high-level features. That means it is inefficient. For addressing this problem, they presented an encoder-decoder architecture. The encoder is a typical pretraining convolution



(a) 1-dilated (b) 2-dilated

(c) 3-dilated

Figure 2: 3×3 expansion convolution, the expansion rate is different: 1, 2, and 3.

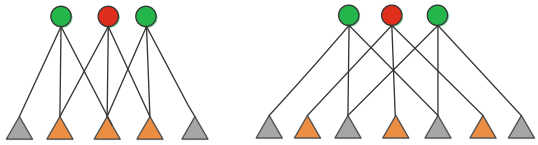


Figure 3: Illustration of the hole algorithm. 3×3 dilated convolutions with rate = 2).

network while a decoder consists of a deconvolutional layer and an upper sampling layer. Down sampling is aimed at capturing semantic or context information, while up sampling is aimed at recovering spatial information. Because the encoder lessens the image resolution, the segmentation has too few well-defined edges, meaning that the boundaries between the images are not clearly defined.

In [8], the final image prediction is usually reduced by 32 times in several stages of stride convolution and spatial pool, resulting in the loss of fine image structure information and inaccurate prediction, especially at the object boundary. DeepLab [12, 14–16] uses atrous (also names dilation) convolution to expand the receptive field while maintaining the high-resolution feature map, or use the encoder-decoder architecture to solve this problem. It regards the backbone network as an encoder and is responsible for encoding the original input image as a low-resolution feature map.

2.4. Atrous Spatial Pyramid Pooling (ASPP). The ASPP module was first proposed in [17] and further revised in [12]. In ASPP module, as shown in Figure 4, different atrous rates are used to extract multiple scale information. In conclusion, one 1×1 convolution block and three 3×3 convolution blocks have different shrinkage rates (6, 12, and 18, respectively), and one GAP block is employed in parallel. ASPP with different sampling rates and multiple views can capture objects at multiple scales.

It can be found that the receptive field has changed from 3 to 5, approximately doubled; the convolution kernel size is still 3×3, and the input stripe is 2, which is now called dilate rate [12, 14].

3. Our Approach

In this part, we introduce our presented network architecture and then explain the formation of each module in detail. We also propose a loss function as a way of further improving the performance of semantic segmentation.

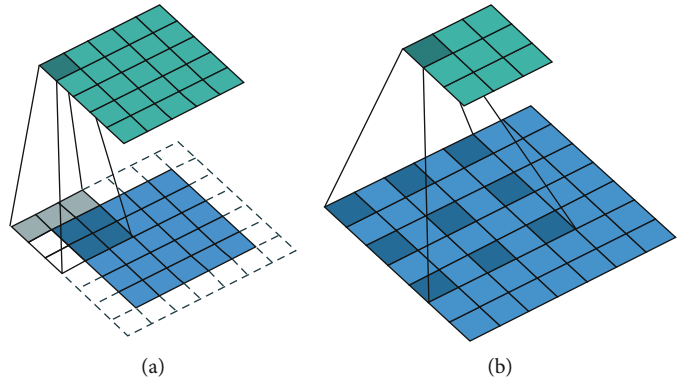


Figure 4: (a) The standard convolution of 3 × 3 kernel. (b) Expansion convolution of 3 × 3 kernel (expansion rate = 2).

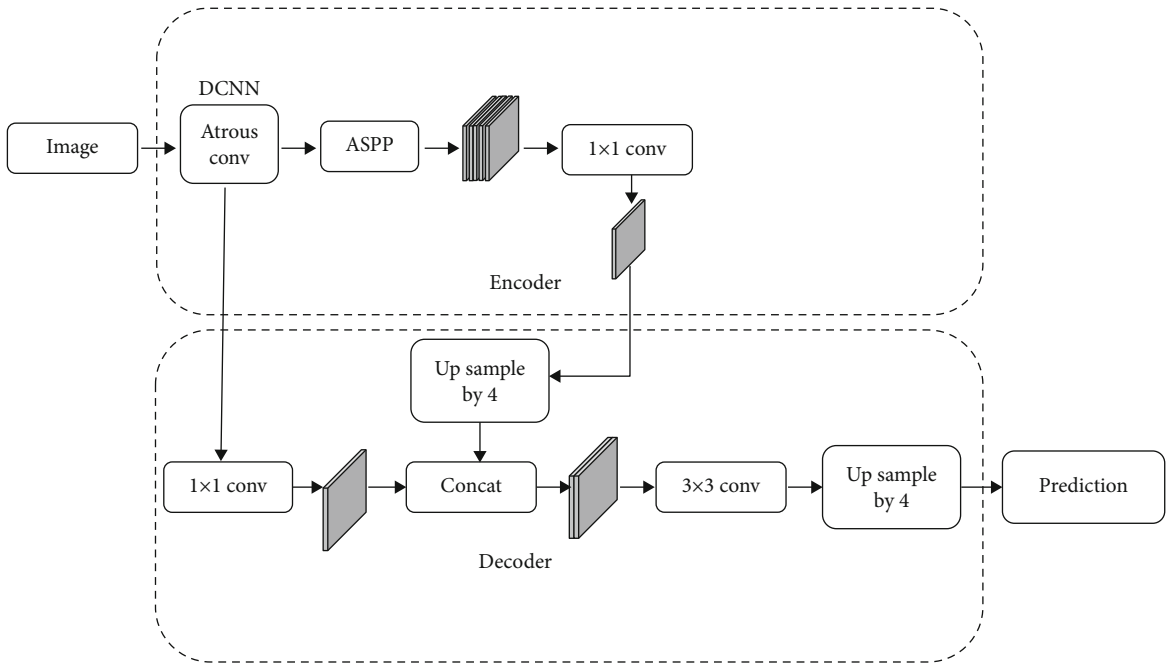


Figure 5: An image segmentation network architecture based on encoder-decoder structure.

3.1. Network Architecture. Figure 5 shows the network architecture including two parts: the encoder is used to extract the feature map and the decoder is used for recovering the resolution of the channel. The amount of parameters in the ASPP part and the decoder part are also huge. Therefore, all the ordinary convolutions are replaced by the depthwise separable convolution. At the same time, the number of channels in ASPP and decoder is also decreased. The backbone network and the ASPP module together constitute the encode module of the network. Input any size of image to obtain the corresponding high-level feature map. Then, through the bilinear up sampling and the low-level feature map of one layer of the encode module, the decode module of the network is formed. Finally, the up sampling is back to the original map size, and the corresponding segmentation map is obtained through the softmax classification layer. This is to decouple spatial information and depth information. It is found that the effect of detail set 1/2 of the size of the feature map and the decoder feature are fused, and finally good results are achieved.

3.2. Backbone Network. Over the past few years, some backbone networks of CNN have achieved great progress in visual missions, showing the most advanced level. It is stacked in the order of convolutional layer, pooling layer, activation function layer, and a fully connected layer. CNN can output the classification score corresponding to the image by inputting the image. In 2012, AlexNet [18] won the title of ILSVRC [19]. AlexNet addresses the problem of image classification and creates a new situation of computer vision. Then, top competitors put forward various CNN architectures, GoogLeNet [8], ResNet [20], DenseNet [21], etc [22]. These network structures can well extract the feature mapping of an image, which lays a solid foundation for semantic

Table 1: Common image classification network infomation summary.

Name	AlexNet	VGG	GoogLeNet	ResNet	Inception	Xception	EfficientNet
Year	2012	2014	2014	2015	2015	2016	2019
Layer	8	19	22	152	/	/	/
Conv	5	16	21	151	/	/	/
Top 5 (error)%	16.4	7.32	6.67	3.57	3.5	5.5	2.9

segmentation [23, 24]. Our network architecture uses Xception to be the feature extractor. Some common classification networks are shown in Table 1 [25]. We came to a conclusion in the experiment. With high calculation complexity, recognition accuracy is allowed to be low; with many parameters, recognition accuracy is allowed to be low. A good network structure design is very important. Different models have different parameter utilization efficiencies.

3.3. Cross-Entropy Loss and Focal Loss. The common loss function of classification problem is cross-entropy loss. It shows the distance between two probability distributions. The closer they are to the cross-entropy, the closer they are. The cross-entropy approach is a novel general method for combinatorial optimization, multipole optimization, and rare event simulation. The standard loss of binary classification is cross-entropy.

Sometimes we will meet the task of image segmentation, which is that the background accounts for a large proportion, but the object accounts for a small proportion of the seriously imbalanced dataset. At this time, we need to carefully use the loss function. The most commonly used loss functions are as follows:

$$\mathrm{CE}(p, y) = -y \log(p) - (1 - y) \log(1 - p), \tag{1}$$

where $y=y_{\text{truth}}$, $p=y_{\text{pred}}$

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & y = 1, \\ -\log(1 - \mathrm{p}) \text{ otherwise.} \end{cases} \tag{2}$$

From the above formula, we can draw a conclusion: when $y = 1$, the larger y' is, the closer it is to y, that is, the more accurate the prediction is, the smaller the loss is. When $y = 0$, the smaller y' is, the closer it is to y, that is, the more accurate the prediction is, the smaller the loss is. The final loss is the sum of $y = 0$ and $y = 1$. This method has one obvious drawback. While the number of positive samples is far less than the negative samples, that is to say, the number of $y = 0$ is far greater than the number of $y = 1$, and its components will dominate the loss function. The model is heavily biased towards the background.

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t). \tag{3}$$

We define p_t:

$$p_t = \begin{cases} \mathrm{y}' & y = 1, \\ 1 - \mathrm{y}' & \text{otherwise,} \end{cases} \tag{4}$$

and rewrite $\mathrm{CE}(p, y) = \mathrm{CE}(p_t) = -\log(p_t)$.

First of all, the proportion of positive and negative samples should be balanced without using negative sample mining and other means. In this paper, we directly multiply a parameter α in front of the CE loss, so that we can easily control the proportion of negative and positive samples.

We get the balanced cross-entropy loss as

$$\mathrm{CE}(p_t) = -\alpha \log(p_t). \tag{5}$$

In practice, α is a decimal between [0, 1]; it is a fixed value and does not participate in training.

Although the above formula can control the weight of positive and negative samples, it cannot control the weight of easy samples and hard samples.

The γ here is called a focusing parameter, $\gamma > = 0$. A modulating factor $(1 - p_t)^{\gamma}$ is called the modulating factor. In practice, we usually add a parameter α in front of the focal loss:

$$\mathrm{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t). \tag{6}$$

In the process of semantic segmentation, there are more categories corresponding to semantic segmentation than the two classification problem in target detection. If the selected parameters λ and γ are not suitable, the cross-entropy loss weight of these pixels will be reduced. Combined with the above analysis, we propose to increase the weight of difficult samples and keep the weight of simple samples almost unchanged. We find that the best results can be obtained by setting $\alpha = 0.5$ and $\gamma = 2$ in our experimental network.

Focal loss was first proposed in the RetinaNet model [26] to solve the imbalance and difficulty of classification in the training process. In practical application, the combination of focal loss and dice loss usually needs to scale them to the same order of magnitude. Use -log to enlarge dice loss and use alpha to reduce focal loss.

4. Experiments and Results

As a way of proving the effectiveness of our presented framework, we evaluated it on the basis of the benchmark dataset (PASCAL VOC 2012) and the latest methods. In the paper,

we report the experimental outcomes of three mainstream semantic segmentation datasets: PASCAL VOC2012, CamVid [27], and Cityscapes [28].

The mean intersection on union (MIoU) is the standard measure of semantic segmentation. The intersection and union ratio of two sets is calculated. In semantic segmentation, the two sets are base truth value and prediction segmentation. This proportion can be morphed to TP (intersection set) over TP, FP, and FN (union set). Calculate the IoU of each class and take the average.

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_j^k p_{ij} + \sum_j^k p_{ij} - p_{ji}}, \tag{7}$$

is equivalent to

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}}, \tag{8}$$

First, calculate the intersection and union ratio of each category, and then get the average. TP is the positive sample that has a correct sort, TN is the positive sample that has a wrong sort. FP is the negative sample of sort error. TP can be understood as the intersection of prediction results and labels, while TP + TN + FP is preunion of test results and labels. The closer the intersection is to the union, the more accurate the segmentation is.

We also used several widely used data augmentation strategies in our training, including 50% probability of horizontal flipping and random scaling of images, scale factor between 0.5 and 2.0 in steps of 0.25, fill and randomly crop the scaled image to 513×513. Finally, with a fine tuning learning rate of $2e$-4 is implemented in the model. When we segment some small target parts, we find that the effect of detail segmentation is very poor. To improve the details, 1/2 of the size of the feature map and the decoder feature are fused, and good results are obtained. In the training, the loss function used is an improved version, focal loss. The results show that the improved focal loss can improve semantic scores. The accuracy of the segmentation and the nonequilibrium of the sample are alleviated.

4.1. PASCAL VOC 2012. PASCAL VOC 2012 includes 20 foreground object classes and one background class, including photos from private collections. There are six indoor classes, seven cars, and seven creatures. The dataset contains 1464 columns, 1449 validation, and 1456 variable size test images. We use 512×512 crops as a way of dividing the learning rate of pretraining weight by 8. All other superparameters are the same as those in [16] experiment. Table 2 shows the performance of our algorithm on VOC 2012, and the detailed results comparison with other methods are displayed in Table 3.

According to the evaluation samples on the test set of PASCAL VOC2012 validation set dataset, we can see that the proposed method is applicable to animals, people, and objects. The edge of equal targets can be segmented carefully, which improves the classification accuracy of the stool, animal, bicycle, and so on. The evaluation of the abovementioned classification index shows that its effect is better than many segmentation methods, as shown in Figure 6. Please note that we do not use CRFs for postprocessing, which can smooth the output, but it is too slow in practice, especially for large-scale images.

Table 2: Performance on PASCAL VOC2012 test set.

Method	MIoU
FCN-8s	62.2
ResSegNet	80.4.7%
RefineNet	84.2%
PSPNet	85.4%
DeepLabv3+	87.8%
Ours	85.6%

4.2. Cityscapes. The Cityscapes dataset is a very large image dataset, which focuses on the semantic understanding of street scene. It contains the road driving images of 50 cities in spring, summer, and autumn. There are 19 classes in the dataset, including good weather and moderate weather, many dynamic objects, different scene layouts, and different backgrounds. We have carried out experiments on 5000 fine-labeled images, which are divided into 2975 training images, 500 verification images, and 1525 test images. The resolving power of all images is 1024×2048. It contains 5000 high-quality pixel level annotations of size 1024×2048 (2975, 500, and 1525 for training, verification, and test sets, respectively) and 2975, 500, and 1525 (training, verification, and test sets separately).

As shown in Figure 7, finally, the method achieves 81.79% MIoU precision on Cityscapes test set on 1024×2048 image. Table 4 shows the performance of our algorithm on Cityscapes 2012 test set.

4.3. CamVid. As a way of further proving the effectiveness and robustness of this method, we also assess its performance on the CamVid dataset. The Cambridge-driving Labeled Video Database (CamVid) is the first video collection with object l class semantic tags. The ground truth labels provided by the database associated each pixel with one of the 32 semantic classes. The CamVid dataset contains images of city road driving scenes. We use 11 classes, including 367 training, 101 verification, and 233 test images. The resolution of all images is 720×960.

We train all models from random initialization and fine tune the pretrained parameters on ImageNet. In the training process, the size of random clipping is 512×512, and the batch size is 16. All other superparameters are the same as PASCAL VOC 2012 experiment. After 30000 iterations on the training set, the model in this paper achieves 77.61% MIoU on the validation set and 69.39% MIoU on the test set.

We can see that the models in this paper can get very accurate semantic segmentation results. Whether it is a small target, or some targets with occlusion and overlap, the method in this paper can accurately segment them.

Table 3: Our highest scoring entry in each column is shown in italic. Results in a performance of 85.6% on PASCAL VOC 2012 test set.

Category	FCN-8s	ResSegNet	RefineNet	PSPNet	DeepLabv3+	Ours
Bicycle	34.2	65.2	73.2	72.7	77.1	*78.2*
Chair	21.4	37.4	43.7	43.1	56.9	*57.1*
Sheep	72.4	85.9	92.9	94.4	92.9	*94.4*
Mean	62.2	80.4	84.2	85.4	87.8	85.6

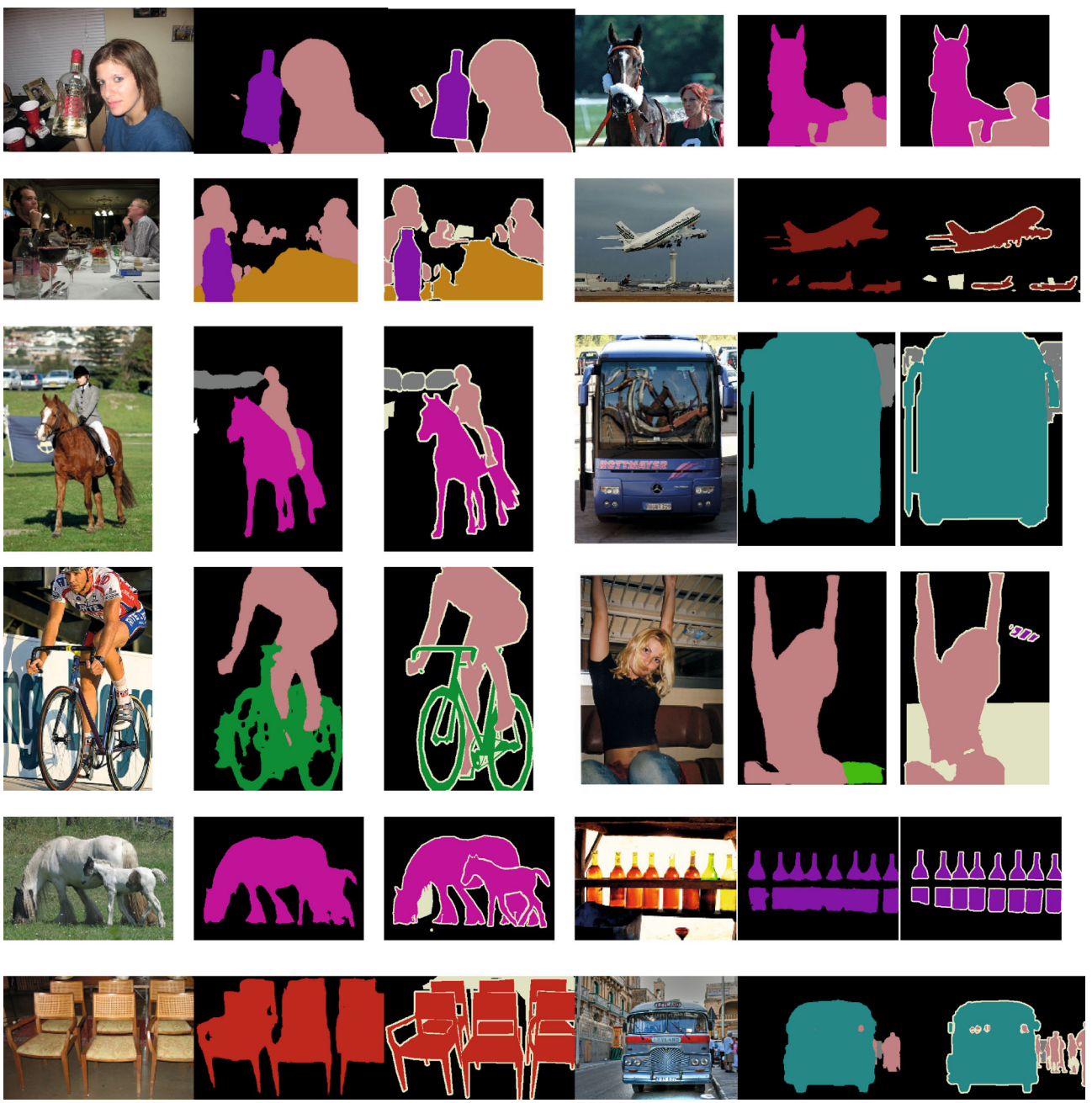


Figure 6: The visualization results on the PASCAL VOC2012 validation set using our methods.

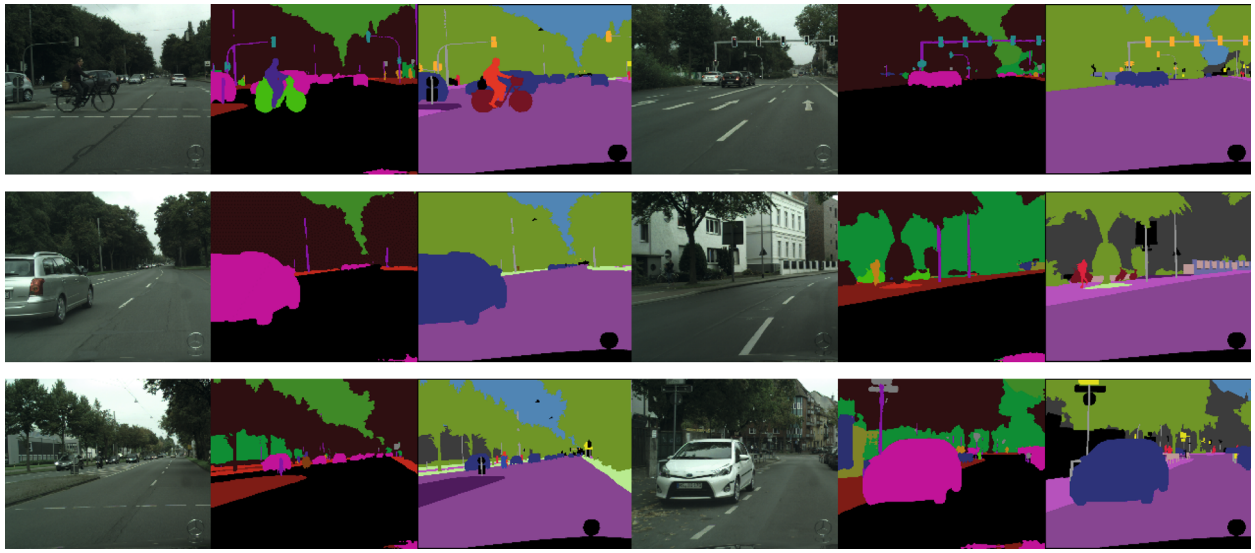


Figure 7: The visualization results on the Cityscapes data using our methods.

Table 4: Segmentation results on Cityscapes test set (https://www.cityscapes-dataset.com/benchmarks/).

Method	MIoU(%)
FCN-8s	65.3
Dilation10	67.1
ShuffleNet v2 + DPC	70.3
MobileNetV2Plus	70.7
ML-CRNN	71.2
Ladder DenseNet	74.3
TuSimple	77.6
DeepLabv3+	82.1
Ours	81.79%

5. Conclusion

We introduce a simpler yet robust network for improving semantic segmentation tasks. Combining ASPP and a classical encoder-decoder structure, an improved loss function more suitable for the application is proposed. The experimental outcomes show the superiority of this method. It not only effectively improves the segmentation performance but also significantly improves the imbalance of training data. As a way of improving the learning ability of this method, we will focus more on weak supervised learning and metalearning down the road. We believe that semantic segmentation can provide a good practice for future smart city construction.

Data Availability

The data used to find the study can be available upon request to the corresponding author.

Conflicts of Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work was supported by the Hubei Natural Science Foundation (2015CFB525), the National Natural Science Foundation (6130329), and the Hubei Natural Science Foundation Innovation Research Group (2017CFA012). We would like to thank those anonymous commentators who helped promote the quality of their papers.

References

[1] D. C. Cirean et al., "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in Neural Information Processing Systems*, vol. 25, pp. 2852–2860, 2012.

[2] M. Oberweger, P. Wohlhart, and V. Lepetit, *Hands deep in deep learning for hand pose estimation*, Computer Vision Winter Workshop (CVWW), 2015.

[3] R. Mottaghi, X. Chen, X. Liu et al., "The role of context for object detection and semantic segmentation in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2013.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012.

[5] B. Luo, Y. Sun, G. Li, D. Chen, and Z. Ju, "Decomposition algorithm for depth image of human health posture based on brain health," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6327–6342, 2020.

[6] D. Jiang, Z. Zheng, G. Li et al., "Gesture recognition based on binocular vision," *Cluster Computing*, vol. 22, Supplement 6, pp. 13261–13271, 2019.

[7] S. Segvic, K. Brkic, Z. Kalafatic, and A. Pinz, "Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle," *Machine Vision and Applications*, vol. 25, no. 3, pp. 649–665, 2014.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[9] Y. Wu and K. He, "Group normalization," in *Computer Vision – ECCV 2018*, Springer, 2018.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[11] M. I. Razzak, S. Naz, and A. Zaib, *Deep learning for medical image processing: overview, challenges and future*, 2017.

[12] L. C. Chen et al., "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *Computer Science*, vol. 4, pp. 357–361, 2014.

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations (ICLR)*, 2016.

[14] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[17] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[18] B. Ma and A. Entezari, "An interactive framework for visualization of weather forecast ensembles," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 1091–1101, 2019.

[19] L. Zhou, C. Zhang, and M. Wu, "D-Linknet: linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, June 2018.

[20] S. Gupta et al., *Learning rich features from RGB-D images for object detection and segmentation*, 2014.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[22] X. Shen, A. Hertzmann, J. Jia et al., "Automatic portrait segmentation for image stylization," *Computer Graphics Forum*, vol. 35, no. 2, pp. 93–102, 2016.

[23] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, no. 1, pp. 64270–64277, 2018.

[24] B. Ma and A. Entezari, "Volumetric feature-based classification and visibility analysis for transfer function design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 12, pp. 3253–3267, 2018.

[25] B. Ma, S. K. Suter, and A. Entezari, "Quality assessment of volume compression approaches using isovalue clustering," *Computers & Graphics*, vol. 63, pp. 18–27, 2017.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[27] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *Computerence*, vol. 39, 2015.

[28] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.