Journal of Theoretical and Applied Information Technology
15th October 2020. Vol.98. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FAST AND ACCURATE INSTANCE SEGMENTATION FOR AUTONOMOUS DRIVING BASED ON REGION-BASED CONVOLUTIONAL NEURAL NETWORK

HOANH NGUYEN
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

This paper presents a two-stage framework for fast and accurate instance segmentation of objects in traffic scene images based on region-based convolutional neural network. For improving the inference speed of the proposed framework, a lightweight deep convolutional neural network which achieved high accuracy in very limited computational budgets is adopted to generate base feature maps. To enhance the segmentation performance on small objects, this paper designs an enhanced module to generate fused feature map which improves the resolution of small objects and simultaneously includes more semantic information. The fused feature map enhances the classification performance and the segmentation performance of small objects. Furthermore, an improved RoI pooling process based on deformable RoI pooling is proposed in this paper. The improved RoI pooling employs a lightweight offset prediction branch which contains fewer parameters compared with standard offset prediction branch, thus improving the inference speed of the proposed framework. For evaluating the proposed framework on instance segmentation of objects in traffic scene images, the Cityscapes dataset is adopted. Experimental results show the effectiveness of the proposed method on both accuracy and inference speed.

Keywords: *Instance Segmentation, Autonomous Driving, RoI Pooling, Deep Convolutional Neural Network, Region-based Convolutional Neural Network*

1. INTRODUCTION

Recently, image segmentation is applied in many problems, such as medical image analysis systems, autonomous driving systems, video surveillance systems, and so on. Image segmentation includes semantic segmentation and instance segmentation. Semantic segmentation can be formulated as a classification of pixels with semantic labels, where pixels are labelled based on a set of fixed object categories. Instance segmentation further extends semantic segmentation problem by separating each of detected object in image. Recently, many methods for image segmentation have been proposed and applied in real world problems. These methods can be divided into two groups: traditional approaches and deep learning approaches. Traditional approaches include region merging [1], k-means clustering [2], clustering techniques [3], conditional and Markov random fields [4], K-nearest neighbor [5], and sparsity-based [6], and so on.

With the fast development of deep learning in recent years, a variety of image segmentation approaches based on deep learning have been proposed. The deep convolution neural networks (CNNs) can learn the features of the objects to be segmented with the dataset autonomously and improve the performance of its model gradually. CNNs mainly consist of three type of layers: convolutional layers, which uses a filter of weights to extract features from image; nonlinear layers, which apply an activation function on feature maps to enable the modeling of non-linear functions by the network; and pooling layers, which replace a small region of a feature map with some statistical information to reduce spatial resolution. Each unit in every layer receives weighted inputs from a small region of units in the previous layer. This small region is called receptive field. In CNNs, the higher-level layers learn features from increasingly wider

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

receptive fields. The main computational advantage of CNNs is that all the receptive fields in a layer share weights, resulting in a significantly smaller number of parameters than fully connected neural networks. Since the development of fully convolutional networks [7], the accuracy of image segmentation has been improved rapidly. Recently, the researcher has been tackling the more challenging instance segmentation task, whose goal is to localize object instances with pixel-level accuracy, jointly solving object detection and semantic segmentation. In this paper, a deep learning-based framework for fast and accurate instance segmentation of objects in traffic scene images is introduced. The proposed framework improves the inference speed on instance segmentation tasks and the accuracy on segmentation of small targets. For improving the inference speed of the proposed framework, a lightweight deep convolutional neural network which achieved high accuracy in very limited computational budgets is adopted to generate base feature maps. For enhancing the segmentation performance on small objects, this paper designs an enhanced module to generate fused feature map which improves the resolution of small objects and simultaneously includes more semantic information. Furthermore, an improved RoI pooling process based on deformable RoI pooling is proposed in this paper. The improved RoI pooling employs a lightweight offset prediction branch which contains fewer parameters compared with standard offset prediction branch, thus improving the inference speed of the proposed framework.

The remaining of this paper is organized as follows. Section 2 introduces the related work. Section 3 details the proposed framework. Section 4 provides the experimental results and comparison between the proposed method and other methods on public datasets. Finally, the conclusions and future works is drawn in Section 5.

2. RELATED WORK

Earlier methods for image segmentation include region merging [1], k-means clustering [2], clustering techniques [3], conditional and Markov random fields [4], K-nearest neighbor [5], and sparsity-based [6]. These traditional methods adopt hand-crafted features for segmenting each pixel in image. Driven by the effectiveness of deep CNNs recently, many methods for image segmentation have been proposed and achieved great improvements. Long et al. [7] proposed to use fully convolutional network (FCN) for semantic segmentation. FCN takes input image of arbitrary size and produces correspondingly sized output with efficient inference and learning. Liu et al. [8] introduced ParseNet, which added global context to deep convolutional networks for semantic segmentation. ParseNet used the average feature for a layer to augment the features at each location. To integrate more context, several approaches incorporate probabilistic graphical models. Chen et al. [9] proposed to combine the responses at the final deep CNN layer with a fully connected Conditional Random Field. The proposed model showed that it is able to localize segment boundaries at a higher accuracy rate than previous methods. Schwing and Urtasun [10] presented a method that jointly trains CNNs and fully connected CRFs for semantic image segmentation. This model achieved encouraging results on the challenging PASCAL VOC 2012 dataset. In [11], the authors proposed to formulate Conditional Random Fields (CRFs) with CNN-based pairwise potential functions to capture semantic correlations between neighboring patches. In addition, an effective network with traditional multi-scale image input and sliding pyramid pooling is designed for improving performance. Another popular branch of deep networks for image segmentation is based on the encoder-decoder architecture. Noh et al. [12] proposed a novel semantic segmentation algorithm by learning a deep deconvolution network. This model mitigates the limitations of the existing methods based on fully convolutional networks by integrating deep deconvolution network and proposal-wise prediction. Badrinarayanan et al. [13] proposed a novel and practical deep fully convolutional neural network architecture for semantic pixel-wise segmentation termed SegNet. The main novelty of SegNet is in the way the decoder upsamples its lower resolution input feature maps. Specifically, it uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to up-sample. In [14], the authors first learned object regions under the supervision of the ground-truth segmentation. The object region representation was then computed by aggregating the representations of the pixels lying in the object region. Finally, the relation between each pixel and each object region were computed, and the representation of each pixel with the object-contextual representation was augmented. Zhang et al. [15] developed a road segmentation/extraction algorithm based on U-Net [16]. U-Net is a well-known architecture. Various extensions of U-Net have been developed for different kinds of images. To integrate the resolution

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

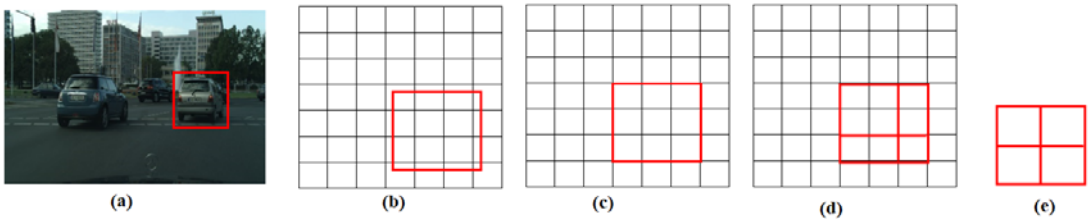

Figure 1: Example of RoI Pooling with 7×7 Input Feature Map. (a) Input Image with Region Proposal Generated by the RPN, (b) The Coordinates of Proposal on The Last Feature Map, (c) The Coordinates of Proposal on The Last Feature Map After The First Step of RoI Pooling, (d) Cropped Proposal Are Divided into Bins to Generate 2×2 Feature Map, (e) Output Feature Map of RoI Pooling After Max Pooling.

and semantic information of different convolution layers, many methods have been designed based on the Feature Pyramid Network (FPN) proposed by Lin et al. [17]. Zhao et al. [18] proposed the Pyramid Scene Parsing Network, which exploits the capability of global context information by different-region-based context aggregation through a pyramid pooling module together with the proposed pyramid scene parsing network. This approach achieved state-of-the-art performance on various datasets. He et al. [19] proposed Adaptive Pyramid Context Network (APCNet) for semantic segmentation. APCNet adaptively constructs multi-scale contextual representations with multiple well-designed Adaptive Context Modules (ACMs). Each ACM leverages a global image representation as a guidance to estimate the local affinity coefficients for each sub-region, and then calculates a context vector with these affinities. Faster R-CNN [20] is a popular two-stage framework, which achieved state-of-the-art performance on object detection. Based on Faster R-CNN, He et al. [21] proposed a Mask R-CNN for object instance segmentation. Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. Liu et al. [22] proposed The Path Aggregation Network (PANet) based on the Mask R-CNN and FPN models. PANet enhances the entire feature hierarchy with accurate localization signals in lower layers by bottom-up path augmentation, which shortens the information path between lower layers and topmost feature. In addition, an adaptive feature pooling which links feature grid and all feature levels to make useful information in each level propagate directly to following proposal subnetworks was introduced. Hu et al. [23] proposed a new partially supervised training paradigm, together with a novel weight transfer function, that enables training instance segmentation models on a large set of categories all of which have box annotations, but only a small fraction of which have mask annotations.

3. PROPOSED METHOD

3.1 Improved RoI Pooling for Enhancing Instance Segmentation

RoI pooling is first introduced in Fast R-CNN [24] for extracting fixed-sized feature maps for each proposal generated by the first stage. Fixed size feature maps are needed for the R-CNN in order to classify them into a fixed number of classes. In RoI pooling process, the features inside any valid region of interest are converted into a small feature map with a fixed spatial by max pooling operation. Although RoI pooling performs well in classification problems, which is robust to small translations, the misalignments between the RoI and the extracted features occur in RoI pooling process have a large negative effect on predicting pixel-accurate masks. In [21], the authors proposed RoIAlign for extracting fixed-sized feature maps for each proposal generated by the region proposal network. RoIAlign removes the harsh quantization of RoI pooling, properly aligning the extracted features with the input. RoIAlign led to a large improvement in instance segmentation compared with RoI pooling. In [25], Deformation RoI pooling is introduced. Deformation RoI pooling includes a RoI pooling process followed by a fully connected layer to generate the normalized offsets, which are then added to the spatial binning positions. After generating offsets, the deformable RoI pooling employs RoI pooling to generate the output feature map based on input regions with augmented offsets.

To elaborate on the improved RoI Pooling proposed in this paper, the RoI pooling, RoIAlign, and Deformation RoI pooling are explained first, and then the improved RoI Pooling is introduced.

3.1.1 RoI pooling

In Faster R-CNN [20], the RPN generates region proposals with the offsets for each anchor based on the last feature map of the base network. The coordinates of proposals, which are presented

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

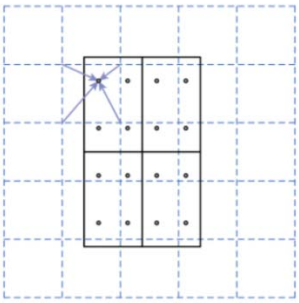

Figure 2: RoIAlign Operation.

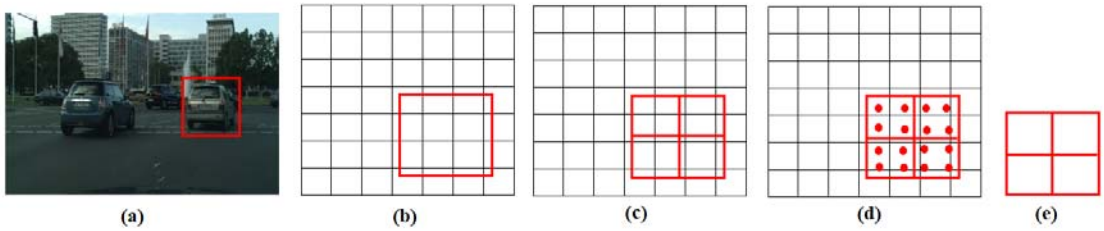

Figure 3: Example of RoIAlign with 7×7 Input Feature Map. (a) Input Image with Region Proposal Generated by the RPN, (b) The Coordinates of Proposal on The Last Feature Map, (c) Cropped Proposal Are Divided into Bins to Generate 2×2 Feature Map, (d) Four Regularly Sampled Locations in Each RoI Bin Are Calculated by Bilinear Interpolation, (e) Output Feature Map of RoIAlign After Max Pooling.

based on original image size, can be obtains based on these offsets. The last feature map was decreased k times from the original image via convolution layers and pooling layers. To get the coordinates of proposals relative to the last feature map size and generate fixed size feature map, RoI pooling first divides each coordinate generated by the RPN by k and take an integer part (e.g., $[x/k]$). Next, new coordinates are used to crop proposal from the last feature map. Then, cropped proposals are divided into bins (e.g., 7×7), and the maximum value in each bin is taken as the value of pixel in the fixed size feature map. Figure 1 shows an example of RoI pooling process with 7×7 input feature map. RoI pooling performs well in box classification. However, the misalignments between the RoI and the extracted features occur in RoI pooling process have a large negative effect on predicting pixel-accurate masks.

3.1.2 RoIAlign

RoIAlign was first introduced in Mask R-CNN [21] to mitigate the misalignments between the RoI and the extracted features in RoI pooling process. In the first step of RoIAlign, each coordinate generated by the RPN is divided by k without rounding. Thus, new coordinates relative to the size of the last feature map are float values. In the second step, cropped region in the last feature map is divided into grid (e.g., 2×2). For generating values in these bins, RoIAlign chooses four regularly sampled locations in each RoI bin and use bilinear interpolation [26] to compute the exact values of the input features at these locations as shown in Figure 2. In the final step, among these four points, maximum or average value from each bin is taken as the value of pixel in the fixed size feature map. Figure 3 shows an example of RoIAlign with 7×7 input feature map. RoIAlign removes the harsh quantization of RoI pooling, properly aligning the extracted features with the input. RoIAlign leads to large improvements in instance segmentation.

3.1.3 Deformable RoI pooling

Deformable RoI pooling is introduced in [25] as shown in Figure 4. In Deformable RoI pooling process, pooled feature map is first generated by adopting regular RoI pooling. From the pooled feature map, a fully connected layer is used to generate the normalized offsets, which are then added to the spatial binning positions. The offset normalization is necessary to make the offset learning invariant to RoI size. After generating offsets, the deformable RoI pooling employs RoI pooling to generate the output feature map based on input regions with augmented offsets.

3.1.4 Improved RoI pooling

Figure 5 illustrates the structure of the improved RoI pooling proposed in this paper. The proposed improved RoI pooling is inspired by

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

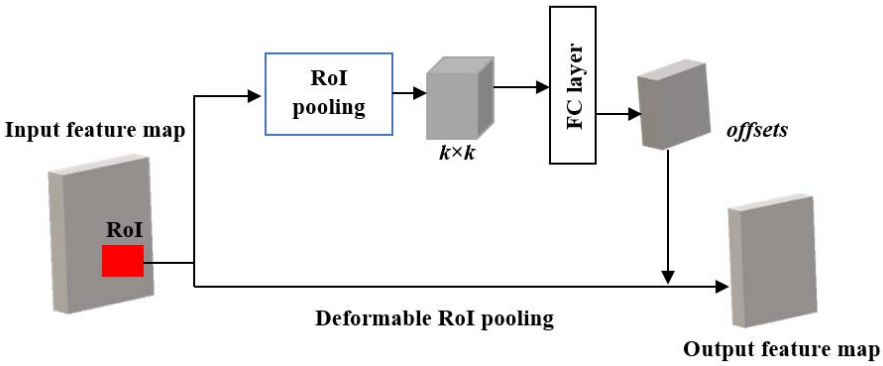

Figure 4: The Structure of Deformable RoI Pooling.

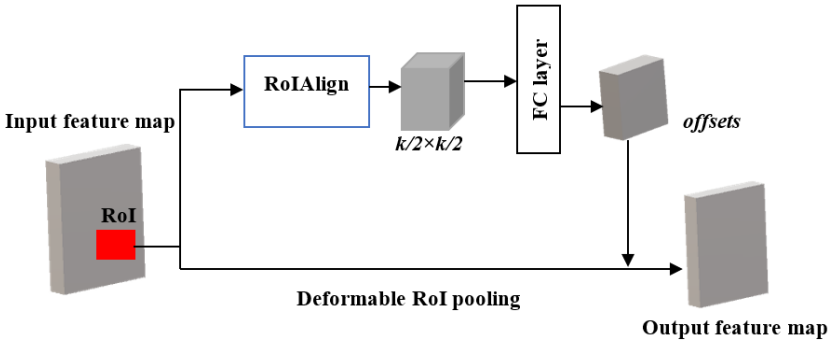

Figure 5: The Structure of The Improved RoI Pooling Proposed in This Paper.

deformable RoI pooling and improves it for instance segmentation in two ways. First, the deformable RoI pooling employs regular RoI pooling process for offset prediction branch, which obtains features from $k \times k$ sub-regions and passes these features through a fully connected layer. Instead, this paper uses a lightweight offset prediction branch which contains fewer parameters than the deformable RoI pooling. More specific, the lightweight offset prediction branch adopts RoIAlign to obtains features from $k/2 \times k/2$ sub-regions followed by a fully connected layer. With smaller input vector of features, the number of parameters in subsequence layer will decrease. Next, the standard deformable RoI pooling employs regular RoI pooling in the fixed size feature map generation branch to generate the output feature map based on input regions with augmented offsets. In contrast, the improved RoI pooling proposed in this paper adopts RoIAlign in the fixed size feature map generation branch to generate the output feature map based on input regions with augmented offsets. As a result, the harsh quantization of RoI pooling is removed, and the extracted features are properly aligning with the input, thus leading to large improvements in instance segmentation as shown in the experimental results section.

3.2 Network Architecture

Figure 6 presents the overall architecture of the proposed network. The proposed network is designed based on two-stage framework. The region proposal network [20] is used at first stage to generate object proposals, and the network head is adopted at second stage for bounding box recognition and mask prediction to each object proposal. For the backbone network, this paper adopts ShuffleNet architecture [27] to generate the base feature maps. ShuffleNet is a lightweight deep CNN network which achieves the best accuracy in very limited computational budgets. By shuffling the channels, ShuffleNet outperformed MobileNetV1 [28]. Figure 7 shows the architecture of ShuffleNet network. There are total four blocks in ShuffleNet followed by a max pooling layer and a dense layer. The first block consists of a convolution layer and a max pooling layer. Other blocks are composed of a stack of ShuffleNet units. The number of bottleneck channels is set to 1/4 of the output channels for each ShuffleNet unit. This paper discards the last max polling layer and dense layer in ShuffleNet architecture.

To enhance the segmentation performance on small objects, this paper designs an enhanced module to generate fused feature map which

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

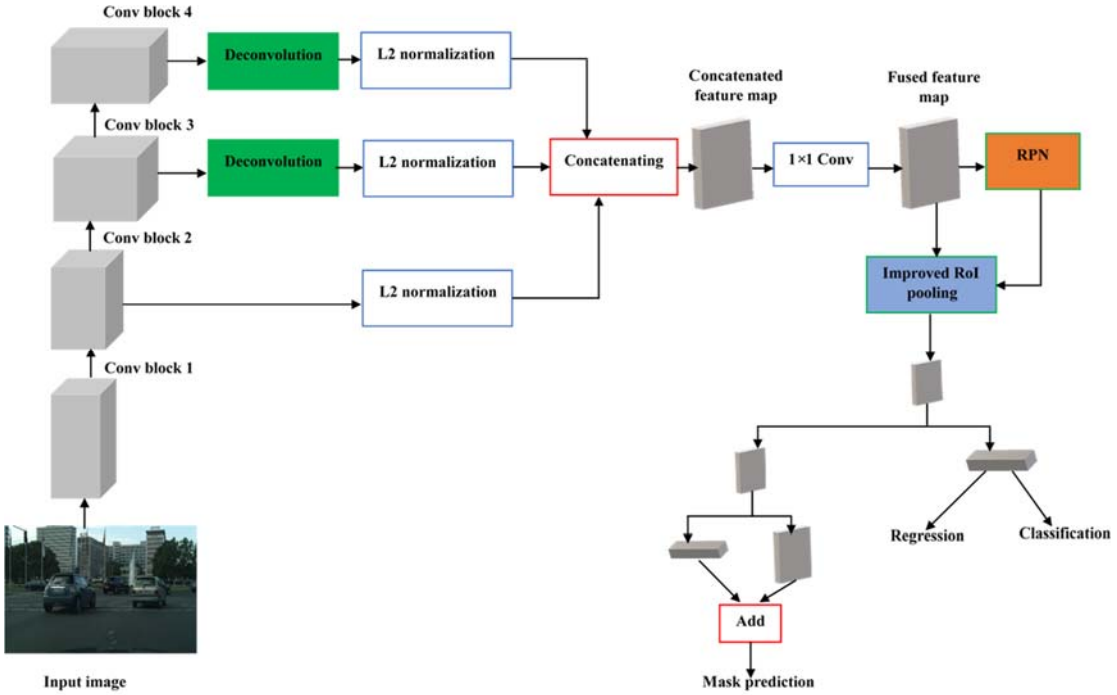

Figure 6: The Overall Architecture of The Proposed Network.

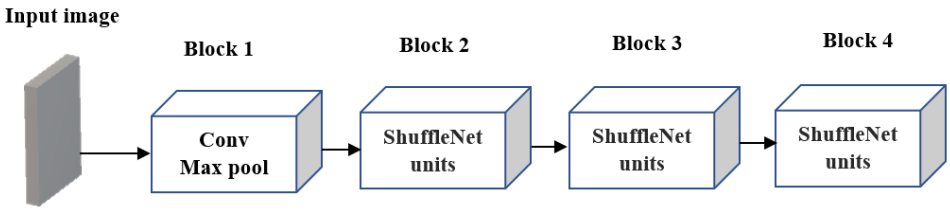

Figure 7: The Architecture of Reduced ShuffleNet Used in This Paper.

improves the resolution of small objects and simultaneously includes more semantic information which enhances the classification performance. In the enhanced module, multi-scale deconvolution operation is first used to upsample the output of the third and fourth convolution block. Next, the feature maps at different blocks, including the second convolution block and two deconvolution blocks, are assembled to generate concatenated feature map. It should be noted that L2 normalization is adopted at each feature map before concatenation operation to effectively keep the feature values from different convolution layers on the same scale. For input vector $x = (x_1, x_2, \dots, x_n)$, L2 normalization is defined as follow:

$$\tilde{x} = \frac{x}{\|x\|_2} = \frac{x}{\sqrt{\sum_{i=1}^{n} |x_i|^2}} \quad (1)$$

where $\tilde{x}$ denotes the normalized vector, $\|x\|_2$ denotes the L2 normalization of x, n denotes the number of channels.

Finally, a 1×1 pointwise convolution is used to compress the number of channels within the concatenated feature to generate fused feature map. The fused feature map is used as input features for the RPN and the network head.

For the network head, this paper follows [22] to design the network head. The fixed size feature map generated by the improved RoI pooling module is fed into two subnetworks. The first subnetwork includes two fully connected layers followed by two parallel fully connected layers for classifying and regressing each of proposed. The second subnetwork includes two paths. The upper path consists of four 3×3×256 consecutive convolutional layers and one deconvolutional layer. The deconvolutional layer is used to upsample feature with a factor of two. The upper path predicts a binary pixel-wise mask for each class. The lower path consists of two 3×3 convolutional layers followed by a fully connected layer to predict a class-agnostic foreground/ background mask. Finally, mask of each class from the upper path and mask of foreground/background

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: The Number of Instances of Each Object Category in The Training Set of The Cityscapes Dataset.

Object category	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle
Number of instances	17,900	1,800	26,900	500	400	200	700	3,700

Table 2: Results on The Cityscapes Dataset.

Methods	Results		
	AP (%)	AP_{50} (%)	Time (ms)
DWT [30]	15.6	30	-
SAIS [31]	17.4	36.7	-
SGN [32]	25	44.9	-
Mask R-CNN [21]	26.2	49.9	330
PANet [22]	31.8	57.1	480
Proposed Method	30.6	56.2	180

prediction from the lower path are added to obtain the final mask prediction.

4. EXPERIMENTAL RESULTS

4.1 Dataset and Metrics

The proposed approach for instance segmentation is evaluated on the Cityscapes dataset [29]. Cityscapes is a large-scale database with a focus on semantic understanding of urban street scenes. This dataset contains a diverse set of stereo video sequences recorded in street scenes from 50 cities, with high quality pixel-level annotation of 5,000 frames in the fine training set, in addition to a set of 20,000 weakly annotated frames in the coarse training set, which is not used in this paper for instance segmentation purpose. The fine training set consists of 2,975 images for training, 500 images for validation, and 1,525 images for testing. All images have the resolution of 2048×1024 pixels. The instance segmentation task involves 8 object categories. Table 1 shows the number of instances of each object category in the training set of the fine training set.

For the evaluation metrics, this paper uses the standard COCO metrics, including mask AP (averaged over IoU thresholds) and AP_{50} (mask AP at an IoU of 0.5), for reporting instance segmentation results on the Cityscapes dataset. Note that mask AP is evaluating using mask IoU.

4.2 Experimental Results on Cityscapes Dataset

This paper reports the instance segmentation results on the Cityscapes dataset as shown in Table 2. Five recent approaches for instance segmentation are used to compare the results and show the effectiveness of the proposed approach, including DWT [30], SAIS [31], SGN [32], Mask R-CNN [21], and PANet [22]. DWT proposed to combine intuitions from the classical watershed transform and modern deep learning to produce an energy map of the image where object instances are unambiguously represented as energy basins. SAIS introduced a novel object segment representation based on the distance transform of the object masks. In addition, the authors designed an object mask network with a new residual-deconvolution architecture that infers such a representation and decodes it into the final binary object mask. SGN employed a sequence of neural networks, each solving a sub-grouping problem of increasing semantic complexity in order to gradually compose objects out of pixels. It should be noted that all methods in Table 2 use only the fine training set of the Cityscapes dataset for training network. As shown in Table 2, the proposed method obtains 30.6% of AP and 56.2% of AP_{50} on the Cityscapes test set. The proposed method outperforms DWT, SAIS, SGN, and Mask R-CNN on both AP and AP_{50}. Specially, compared with Mask R-CNN, the proposed method improves the AP and the AP_{50} by 4.4% and 6.3% respectively. It can be seen that PANet achieves the best results on the Cityscapes dataset with 31.8% of AP and 57.1% of AP_{50}. However, PANet is slower than the proposed method with 0.48 second for processing an image, while the proposed method takes only 0.18 second. This result shows the effectiveness of the proposed method on both accuracy and inference

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

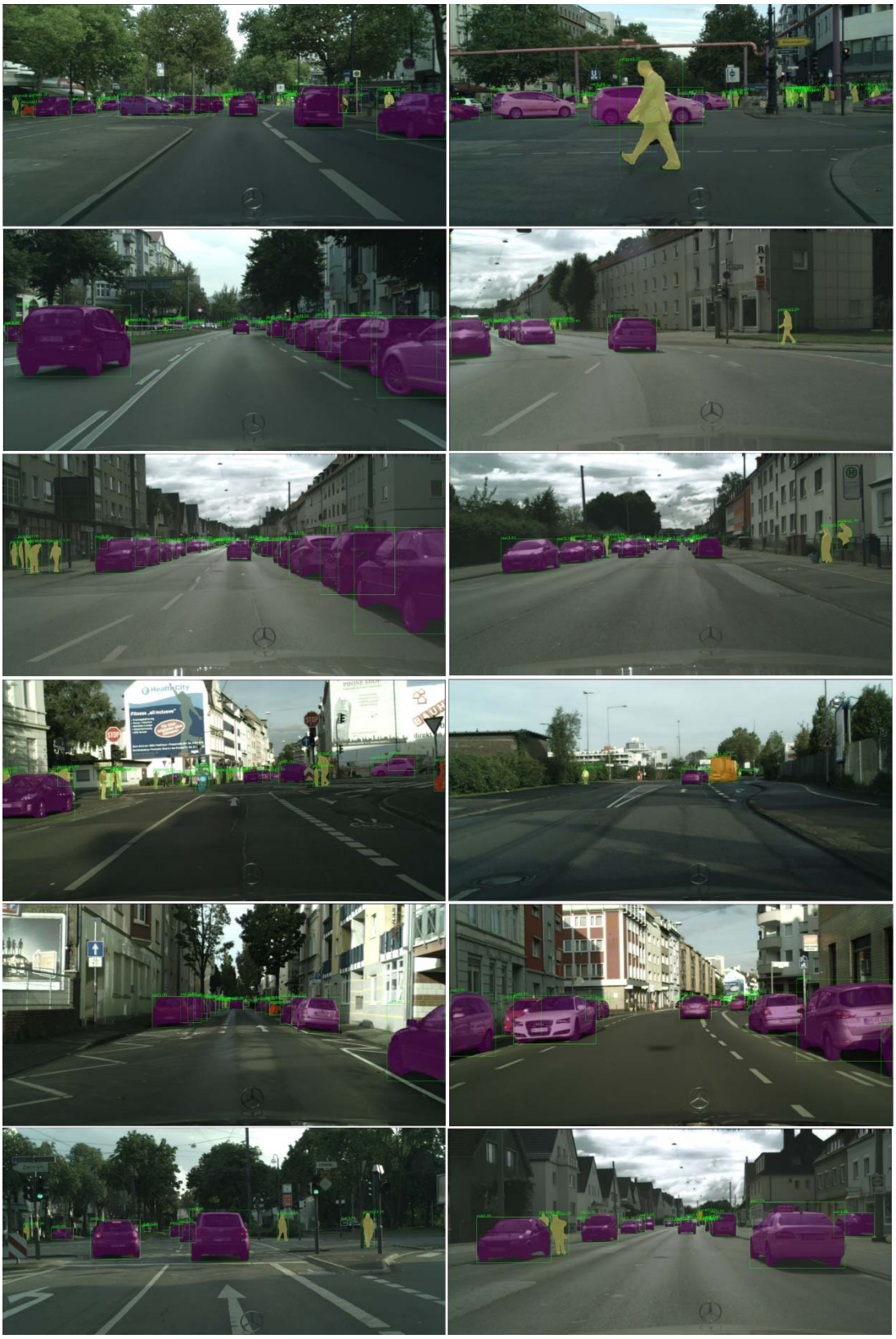

Figure 8: Visual Results of The Proposed Method on The Cityscapes Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

speed. Figure 8 shows some visual results of the proposed method on the Cityscapes dataset.

5. CONCLUSIONS

This paper presents a deep learning-based framework for fast and accurate instance segmentation of objects in traffic scene images. The proposed framework improves the inference speed on instance segmentation tasks and the accuracy on segmentation of small targets. For improving the inference speed of the proposed framework, a lightweight deep convolutional neural network which achieved high accuracy in very limited computational budgets is adopted to generate base feature maps. For enhancing the segmentation performance on small objects, this paper designs an enhanced module to generate fused feature map which improves the resolution of small objects and simultaneously includes more semantic information. Furthermore, an improved RoI pooling process based on deformable RoI pooling is proposed in this paper. The improved RoI pooling employs a lightweight offset prediction branch which contains fewer parameters compared with standard offset prediction branch, thus improving the inference speed of the proposed framework. Experimental results on the Cityscapes dataset show the effectiveness of the proposed method on both accuracy and inference speed.

REFERENCES:

[1] Nock, Richard, and Frank Nielsen. "Statistical region merging." *IEEE Transactions on pattern analysis and machine intelligence* 26, no. 11 (2004): 1452-1458.

[2] Dhanachandra, Nameirakpam, Khumanthem Manglem, and Yambem Jina Chanu. "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm." *Procedia Computer Science* 54 (2015): 764-771.

[3] JABAR, FARAH HA, WAIDAH ISMAIL, ROSALINA A. SALAM, and ROSLINE HASSAN. "IMAGE SEGMENTATION USING A HYBRID CLUSTERING TECHNIQUE AND MEAN SHIFT FOR AUTOMATED DETECTION ACUTE LEUKAEMIA BLOOD CELLS IMAGES." *Journal of Theoretical & Applied Information Technology* 76, no. 1 (2015).

[4] Plath, Nils, Marc Toussaint, and Shinichi Nakajima. "Multi-class image segmentation using conditional random fields and global classification." In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 817-824. 2009.

[5] Kurumalla, Suresh, and P. Srinivasa Rao. "K-nearest neighbor based dbscan clustering algorithm for image segmentation." *Journal of Theoretical and Applied Information Technology* 92, no. 2 (2016): 395.

[6] Minaee, Shervin, and Yao Wang. "An ADMM approach to masked signal decomposition using subspace representation." *IEEE Transactions on Image Processing* 28, no. 7 (2019): 3192-3204.

[7] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[8] Liu, Wei, Andrew Rabinovich, and Alexander C. Berg. "Parsenet: Looking wider to see better." *arXiv preprint arXiv:1506.04579* (2015).

[9] Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Semantic image segmentation with deep convolutional nets and fully connected crfs." *arXiv preprint arXiv:1412.7062* (2014).

[10] Schwing, Alexander G., and Raquel Urtasun. "Fully connected deep structured networks." *arXiv preprint arXiv:1503.02351* (2015).

[11] Lin, Guosheng, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. "Efficient piecewise training of deep structured models for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3194-3203. 2016.

[12] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." In *Proceedings of the IEEE international conference on computer vision*, pp. 1520-1528. 2015.

[13] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 12 (2017): 2481-2495.

[14] Yuan, Yuhui, Xilin Chen, and Jingdong Wang. "Object-contextual representations for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

semantic segmentation." *arXiv preprint arXiv:1909.11065* (2019).

[15] Zhang, Zhengxin, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net." *IEEE Geoscience and Remote Sensing Letters* 15, no. 5 (2018): 749-753.

[16] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241. Springer, Cham, 2015.

[17] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[18] Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881-2890. 2017.

[19] He, Junjun, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. "Adaptive pyramid context network for semantic segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7519-7528. 2019.

[20] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Advances in neural information processing systems*, pp. 91-99. 2015.

[21] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.

[22] Liu, Shu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. "Path aggregation network for instance segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759-8768. 2018.

[23] Hu, Ronghang, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. "Learning to segment every thing." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4233-4241. 2018.

[24] Girshick, Ross. "Fast r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 1440-1448. 2015.

[25] Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. "Deformable convolutional networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 764-773. 2017.

[26] Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. "Spatial transformer networks." In *Advances in neural information processing systems*, pp. 2017-2025. 2015.

[27] Zhang, Xiangyu, Xinyu Zhou, Mengxiao Lin, and Jian Sun. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848-6856. 2018.

[28] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

[29] Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. "The cityscapes dataset for semantic urban scene understanding." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213-3223. 2016.

[30] Bai, Min, and Raquel Urtasun. "Deep watershed transform for instance segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5221-5229. 2017.

[31] Hayder, Zeeshan, Xuming He, and Mathieu Salzmann. "Shape-aware instance segmentation." *arXiv preprint arXiv:1612.03129* 2, no. 5 (2016): 7.

[32] Liu, Shu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. "Sgn: Sequential grouping networks for instance segmentation." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3496-3504. 2017.