Real-time Semantic Segmentation via Region and Pixel Context Network

Yajun Li, Yazhou Liu, Quansen Sun
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China
E-mail: {118106021936, yazhouliu}@njust.edu.cn

1.Motivation

Problems

Full precision semantic segmentation model can achieve great performance, but it will take too much time to calculate. PSPNet takes 1288ms to process one picture, Deeplab v3+ takes even longer (2800ms).

For real-time sematic segmentation, it is necessary to accelerate the inference speed without sacrificing too much quality

Motivation

From a macro perspective, the semantic segmentation task can be divided into two parts: region semantic prediction and pixel-level detail recovery.

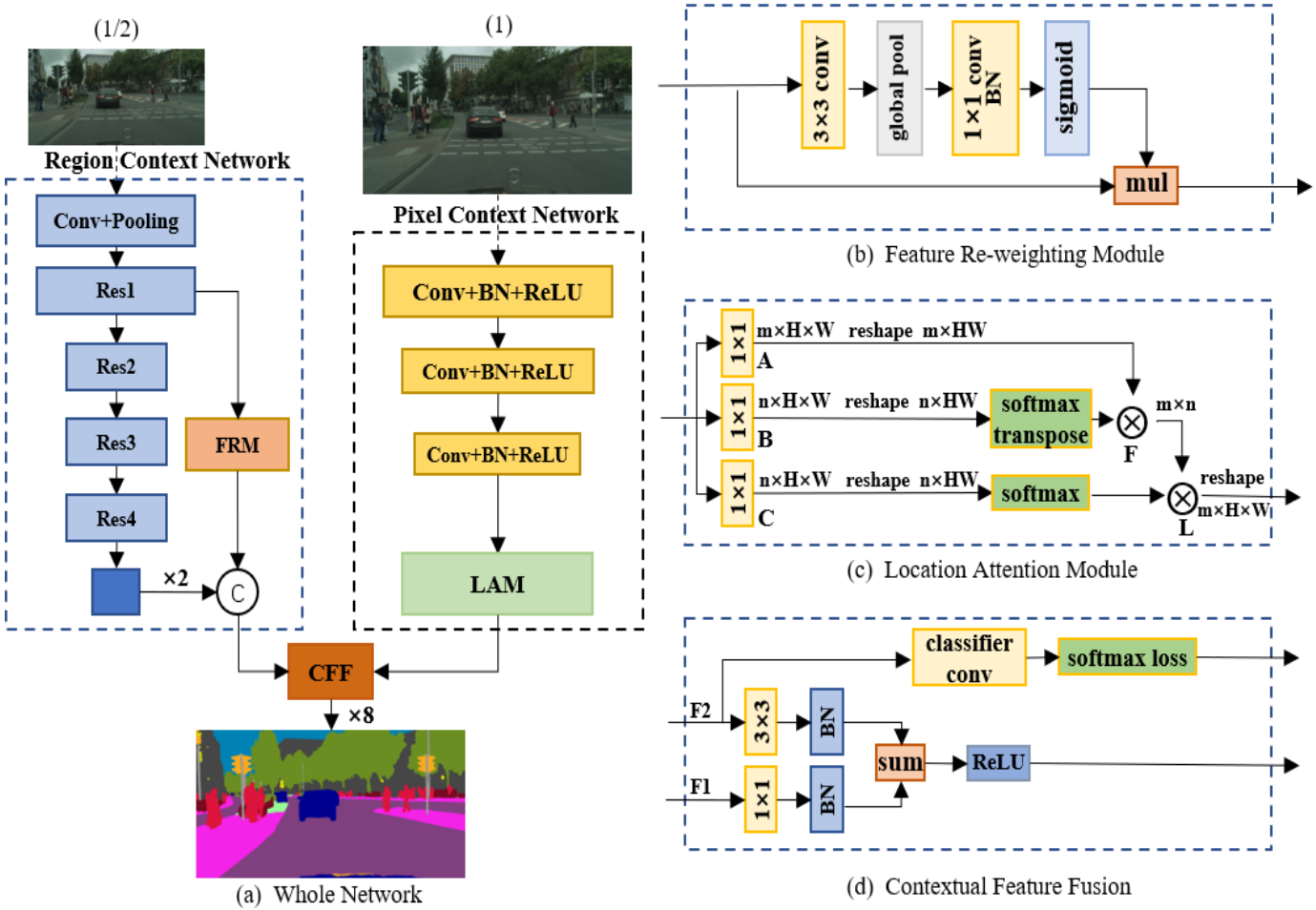
2.Contributions

- We propose a novel Dual Context Network with two sub-networks: Region Context Network and Pixel Context Network. These two sub-networks take different resolution images to accomplish semantic prediction and detail information recovery respectively.

- Feature Re-weighting Module is designed to integrate more context information and Location Attention Module is designed to model spatial interdependencies. In addition, we present Contextual Feature Fusion to further improve the accuracy.

- Experiments prove superior performance of our method through comparison with a number of state-of-the-art networks on the benchmarks of Cityscapes and CamVid.

3. Architecture



4. Ablation study

Method	mIoU(%)
RCN(ResNet50)	67.4
RCN(ResNet50)+FRM	69.1
RCN(ResNet50)+FRM+PCN	70.5
RCN(ResNet50)+FRM+PCN(LAM)	72.9
RCN(ResNet50)+FRM+PCN(LAM)+CFF	74.2
RCN(ResNet18)	62.4
RCN(ResNet18)+FRM	63.7
RCN(ResNet18)+FRM+PCN	65.4
RCN(ResNet18)+FRM+PCN(LAM)	68.6
RCN(ResNet18)+FRM+PCN(LAM)+CFF	69.7

5.Speed and Accuracy Comparisons

Performance on Cityscapes test dataset

Method	Input Size	Time(ms)	Frame(fps)	mIoU(%)
SegNet	640 × 360	16	16.7	57
ENet	640 × 360	7	135.4	57
ESPNet	1024 × 512	9	112	60.3
ICNet	1024 × 2048	33	30.3	69.5
TwoColumn	512 × 1024	68	14.7	72.9
BiSeNet1	768 × 1536	13	72.3	68.4
BiSeNet2	768 ×1536	21	45.7	74.7
DFANet A	1024 × 1024	10	100	71.3
DFANet B	1024 × 1024	8	120	67.1
SwiftNet	1024 × 2048	25	39.9	75.5
Ours(Res50)	512 × 1024 (1024 × 2048)	12	82	**76.1**
Ours(Res18)	512 × 1024 (1024 × 2048)	7	**142**	71.2

Performance on CamVid test dataset

Method	Frame(fps)	mIoU(%)
SegNet	46	46.4
ICNet	27.8	67.1
ENet	-	51.3
BiSeNet1	-	65.6
BiSeNet2	-	68.7
DFANet A	120	64.7
DFANet B	160	59.3
SwiftNet	-	73.86
Ours(Res50)	91	70.8
Ours(Res18)	**166**	66.2

Thank you !