

Flow2Seg: Motion Aided Semantic Segmentation



Xiangtai Li, Jiangang Bai, Kuiyuan Yang, Yunhai Tong(*)

Peking University, Beijing, China

DeepMotion, Beijing, China



Introduction

Semantic segmentation is a fundamental task in computer vision, which aims to predict a semantic category for each pixel in an image. In deep learning era, this field has made steady progress after the Fully Convolutional Networks (FCN). However, most existing methods only take a static image as input and ignores the rich motion information in image sequences.

Motion is an important clue for segmentation task and can separate different objects apart based their different motion patterns, which is complementary to static patterns in an image. Motivated by this, we propose to add one path network named **Flow2Seg** by **taking optical flow as input**, in addition to the image path by one state-of-the-art network(PSPNet).

Contributions

- We propose a novel and light module Flow2Seg for directly mapping optical flow into segmentation map and we explore the usage of FCNs for learning from the optical flow.
- Combined with image segmentation model, we achieve considerable improvement compared with the PSP-net baseline on Cityscapes dataset. We achieves the state-of-art results among the video semantic segmentation.

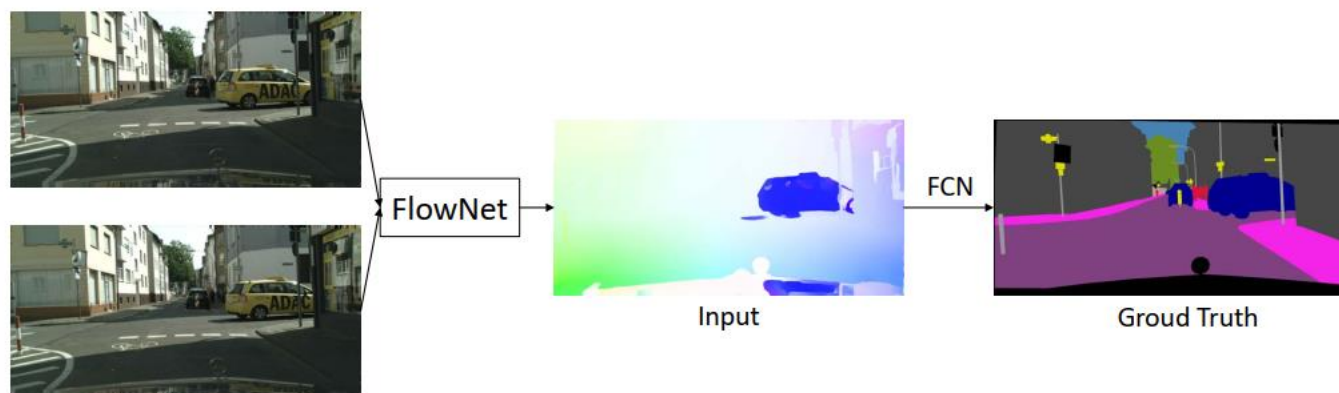


Fig. 1. Overview of Flow2Seg path. Two consecutive frames are used to estimated the optical flow, then the optical flow is fed into FCN for semantic segmentation

Architecture of the Network

• Overview:

Mainly contains Three components: **Flow2Seg**, **Image2Seg** and **Residual Fusion**

- **Flow2Seg:** Unlike very deep networks those used to extract features for image, **relatively shallow ResNet18** is used to process optical flow.
- **Image2Seg:** We choose the previous state-of-the-art model PSPNet as our Image2Seg Module. We use the pretrained ResNet101 with the same dilated strategy as our backbone to extract the feature map.
- **Residual Fusion:**
 - Simply fusing by adding or concatenating their outputs cannot obtain better results since Flow2Seg performs much worse results than Image2Seg.
 - Both output maps from Flow2Seg and Image2Seg are concatenated together followed by two blocks consist of convolution and batch normalization, and a residual fused semantic map is generated with 1×1 convolution.

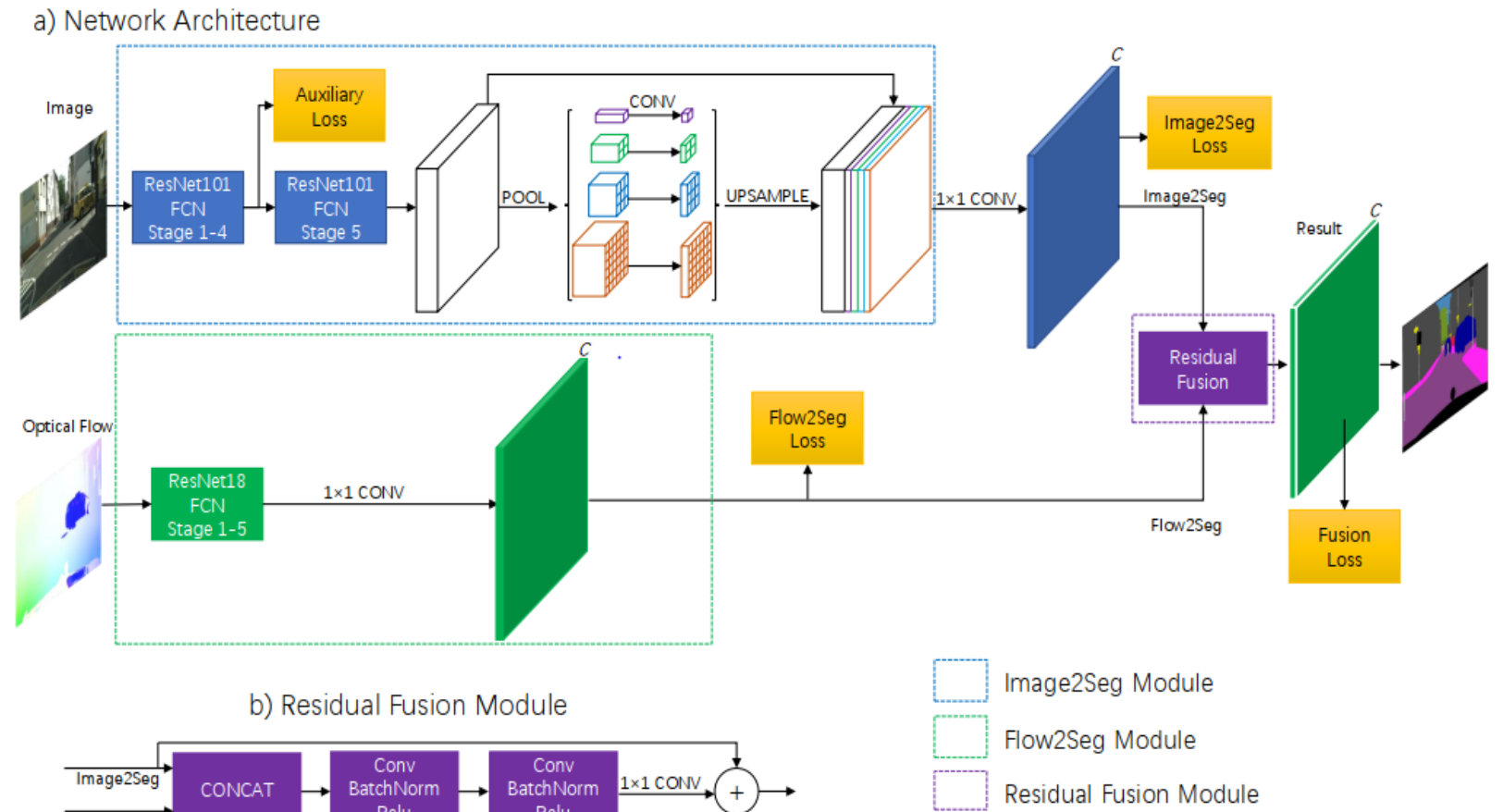


Fig.2. Network Architecture. It contains three different parts: Image2Seg Module, Flow2Seg Module and Residual Fusion Module. (b) Residual Fusion Module. C denotes the number of categories.

• Loss Function

➤ Multi-Task Learning Loss

$$\mathcal{L} = \mathcal{L}_{Image2Seg} + \mathcal{L}_{Fusion} + \alpha * \mathcal{L}_{Flow2Seg} + \beta * \mathcal{L}_{Aux}$$

Experiment Results

Method	mIoU(%)	Method	mIoU(%)
FlowNetS	35.6	ResNet18-FCN	39.6
FlowNet2	39.6	ResNet50-FCN	37.4
PWC	36.3	ResNet101-FCN	35.4
GF-flow	25.4	ResNet18 + ASPP	39.4
		ResNet18 + DenseASPP	39.3
		ResNet18 + PSP	38.2

Tab. 1. Ablation study of Flo2seg Module. (a) different optical flow inputs with resnet18-fcn as backbone. (b) Different backbone networks. (c) Add it to the PSPNet

Method	Backbone	use optical flow	mIoU(%)
RefineNet [22]‡	ResNet101	no	73.6
SAC [40]‡	ResNet101	no	78.1
DUC-HDC [32]‡	ResNet101	no	77.6
AAF [17]‡	ResNet101	no	79.1
BiSeNet [37]‡	ResNet101	no	78.9
PSANet [43]‡	ResNet101	no	80.1
DFN [38]‡	ResNet101	no	79.3
DSSPN [20]‡	ResNet101	no	77.8
Ours‡	ResNet101	yes	80.4

Tab. 2. Compared with state-of-the-art. (a) with Image based models (b) with Video based models

Visualization Results

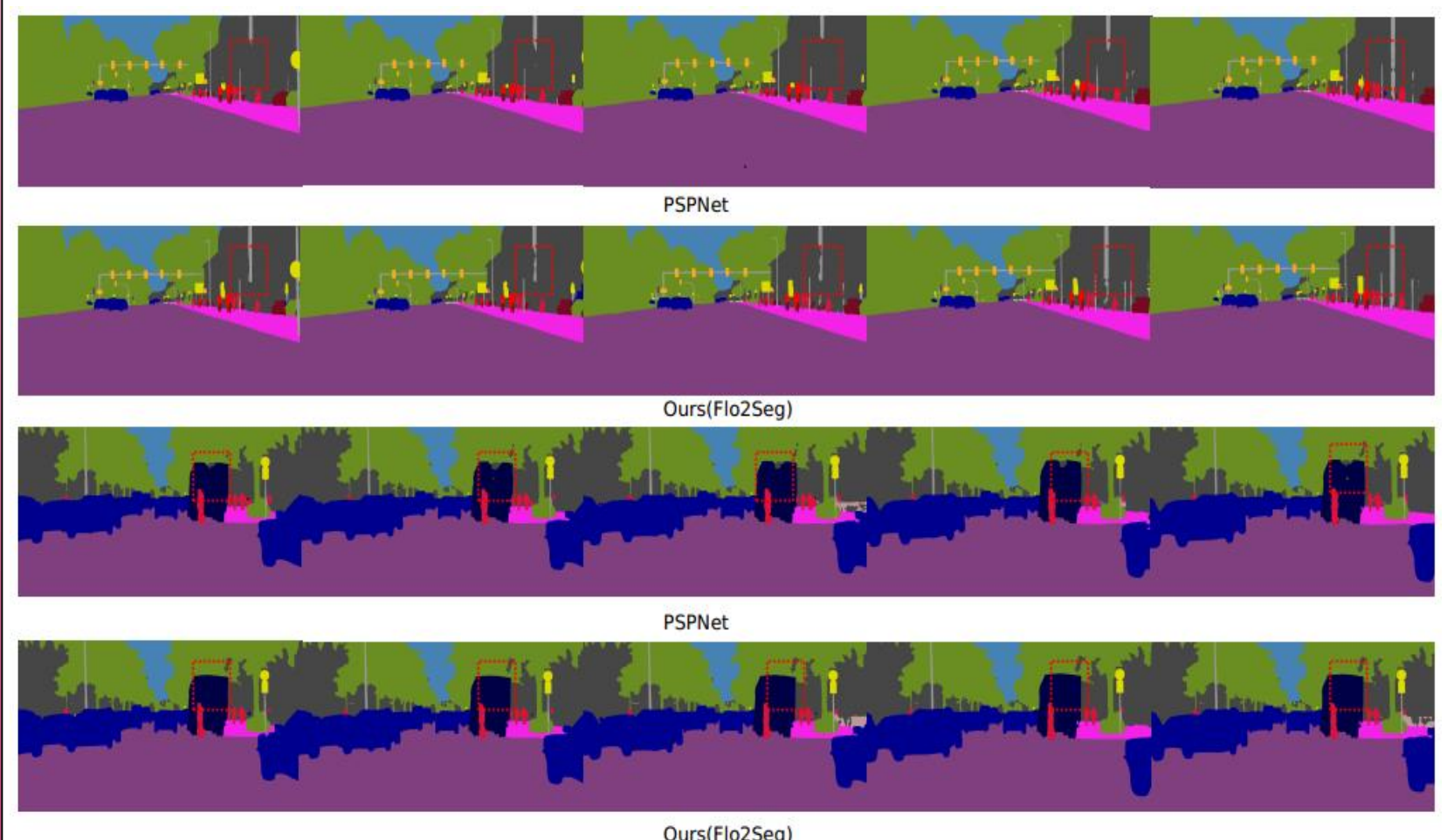


Fig. 3. Comparison of segmentation results of PSPNet and our results on Cityscapes video sequences. First two rows show our method handles missing small objects on successive frames while the last two rows show our method can remove ambiguities of the same object. Both are shown in red boxes

Acknowledge

This work is done when Xiangtai Li was an intern in DeepMotion AI research.

Contact

Email: lxtpku@pku.edu.cn

Wechat: lxtbupt94