A BOUNDARY-AWARE DISTILLATION NETWORK FOR COMPRESSED VIDEO SEMANTIC SEGMENTATION

*Hongchao Lu, Zhidong Deng**
Deparment of Computer Science, Tsinghua University, Beijing 100084, China
luhc15@mails.tsinghua.edu.cn, michael@tsinghua.edu.cn

INTRODUCTION

In this paper, we propose a boundary-aware distillation network (BDNet) in Fig.1 to accelerate the video segmentation as well as improve the segmentation accuracy. A boundary-aware stream is added to perceive the boundary of objects, so as to constrain and guide the features into clear shape. In addition, an auxiliary teacher network is well pretrained on labeled frame so that the main stream enables to learn high-quality knowledge to correct the tailing effect. Our BDNet shows almost 10% time savings as well as 1.6% accuracy improvement over baseline on the Cityscapes dataset.

METHOD

A. Network Architecture

The proposed boundary-aware distillation network (BDNet) is shown in Fig.2, which includes three components:

- The principal part in the middle is the main stream, which indicates a base single-frame semantic segmentation work.
- The boundary-aware stream at the top predicts silhouette of both foreground and background object.
- The distillation stream at the bottom is designed to transfer knowledge from the teacher network T to the main stream.

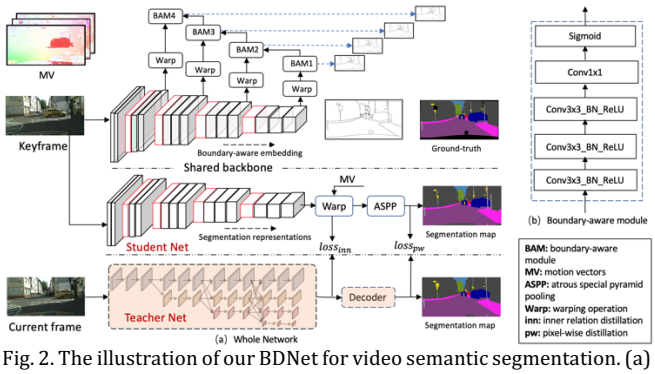

Fig. 2. The illustration of our BDNet for video semantic segmentation. (a) The top part is the boundary-aware stream. The middle one is the main stream, also known as the student network that shares the backbone with the boundary-aware stream. The bottom one is the well-trained teacher network. (b) The structure of boundary-aware module.

EXPERIMENTS

Experiments are conducted on Cityscapes. Two metrics mIoU and s/frame are provided to assess accuracy and speed.

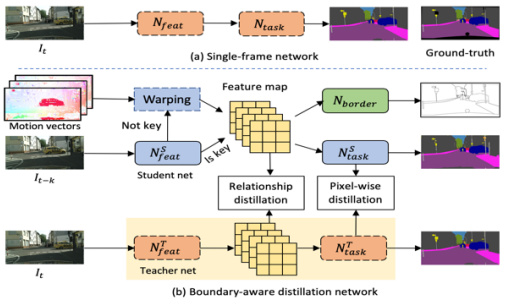

Fig.1. Illustration of (a) single-frame segmentation network and (b) boundary-aware distillation network for video semantic segmentation.

The results are shown as below.

Table 1. The accuracy and inference speed at interval 5 on Cityscapes validation dataset. DL-X is the single frame network DeepLabV2 based on different backbone.

model	accuracy (mIoU, %)	Time (s/frame)
Teacher Net (HRNet)	79.2	-
DFF [5]	68.6	0.32
GRFP [6]	69.4	0.47
DL-18	64.4	0.22
DL-50	69.7	0.48
DL-101	71.3	0.72
Ours BDNet-18	65.4	0.14
Ours BDNet-50	69.7	0.21
Ours BDNet-101	70.2	0.29

Table 2. Comparison based on different backbone on validation dataset. The accuracy is calculated at interval 5.

model	backbone (mIoU, %) DL-101	DL-50	DL-18
single-frame network	71.1	69.7	64.4
+FlowNet(DFF)	68.6	67.4	62.5
+MV	69.0	68.4	63.0
+MV+BAS	69.5	69.1	63.5
+MV+BAS +D_{pw}	69.7	69.4	64.1
+MV+BAS+D_{pw}+D_{inn}	70.2	69.7	65.4

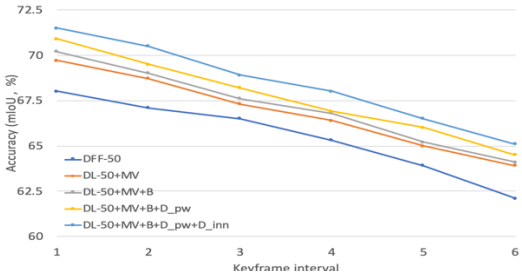

Fig. 3. Accuracy on various keyframe intervals for Cityscapes dataset. The model is based on ResNet-50.