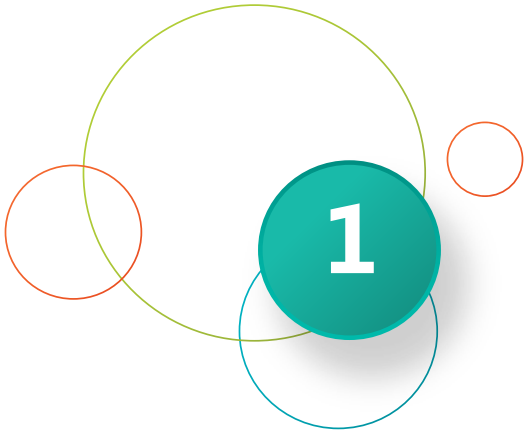
A Boundary-aware Distillation Network for Compressed Video Semantic Segmentation

Hongchao Lu, Zhidong Deng

Deparment of Computer Science,
Tsinghua University, Beijing 100084, China

Contents



Introduction

——Introduction

Video semantic segmentation aims to segment each frame in a video precisely. Both segmentation accuracy and efficiency are concerned about to meet the requirements in applications like self-driving car.

Fig.1 Video semantic segmentation in self-driving task.

Challenges:

- Segmenting each frame in a video is expensive as frame rate and spatial resolution are scaled, ignoring the temporal relationships hidden in neighboring frames.

- Some works utilized optical flow to reuse features as a means to reduce computation cost, introducing additional network to compute optical flow.

——Introduction

The model replaces troublesome optical flow network with block motion vectors encoded in compressed videos (*e.g. H264 codecs*), resulting in shrunken main stream to extract dense features.

The boundary-aware stream perceives the boundary of objects to constrain and guide the features into clear shape.

The teacher network is well pretrained on sparsely labeled frame, so that the main stream enables to learn high-quality knowledge to correct the tailing effect.

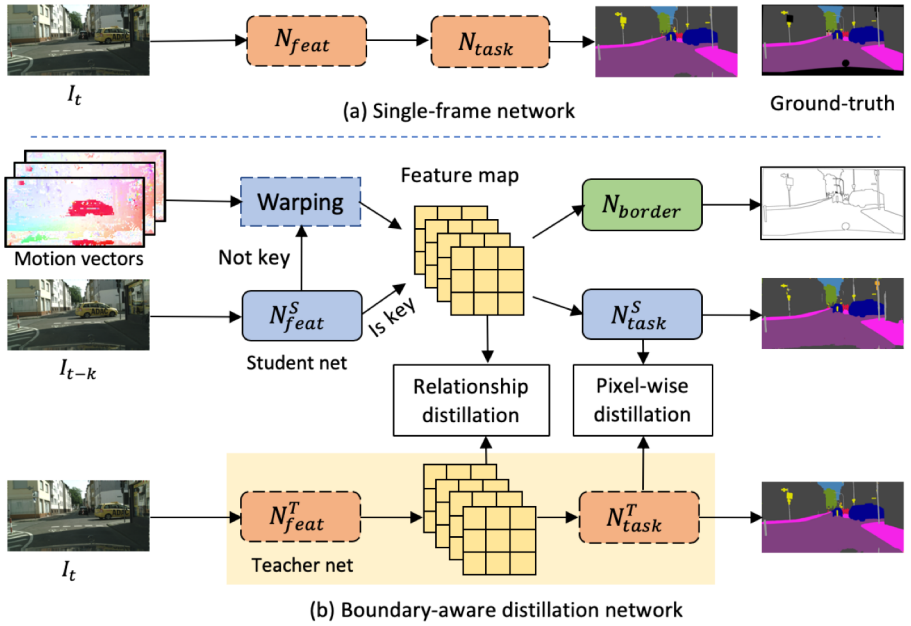
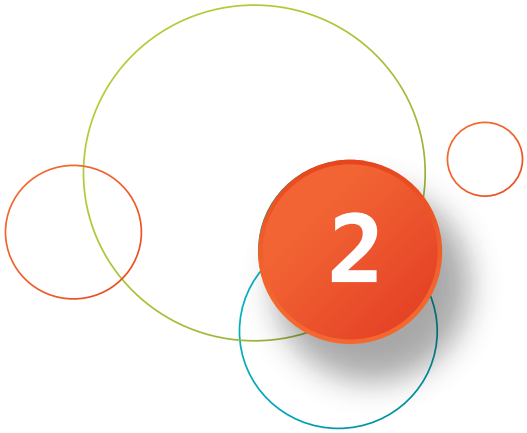


Fig. 1. Illustration of (a) single-frame segmentation network and (b) boundary-aware distillation network (BDNet) for video semantic segmentation.

——Introduction

Contributions:

- We replace optical flow network with block motion vectors embedded in compressed video, which are extracted with negligible computational cost, so as to accelerate video segmentation.
- An auxiliary boundary-aware stream are provided to guide features for discriminative boundary.
- We employ a well pretrained teacher network to transfer knowledge to the main stream for improvement in video segmentation accuracy.



Related Work

Related work:

A. Image Semantic Segmentation

Based on deep convolutional neural networks (DCNN), many approaches achieve the state-of-the-art performance on image semantic segmentation, such as FCN [1], PSP [2], DeepLab [3-5]…

B. Video Semantic Segmentation

DFF [6] [7] shared the features of selected frame and warped then forward with optical flow to save time cost. DVSNet [8] split each frame into four regions and propagated regions based on confidence score.

C. Boundary detection

Boundary detection [9] [10] approximates silhouette of objects, refining the representation areas. In this work, we apply boundary detection in video semantic segmentation.

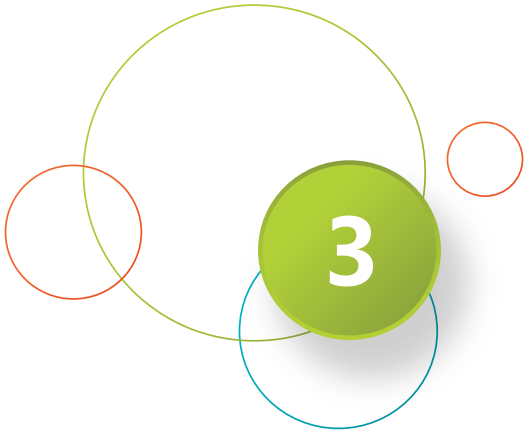
——Related Work

D. Knowledge distillation

Knowledge distillation [11] is developed to transfer useful information from the cumbersome teacher model to a compact student model.

E. Compressed video

Compressed video [12] [13] consists of a group of pictures (GOP), which contains I-frame (intra-coded frame) and P/B-frame (predictive/bi-directional frame). I-frame is generally viewed as RGB image. P-frame stores motion clues representing coarse movement of blocks between current and preceding frame.



Methods

——Methods

Network architecture

The proposed boundary-aware distillation network (BDNet) includes three components:

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

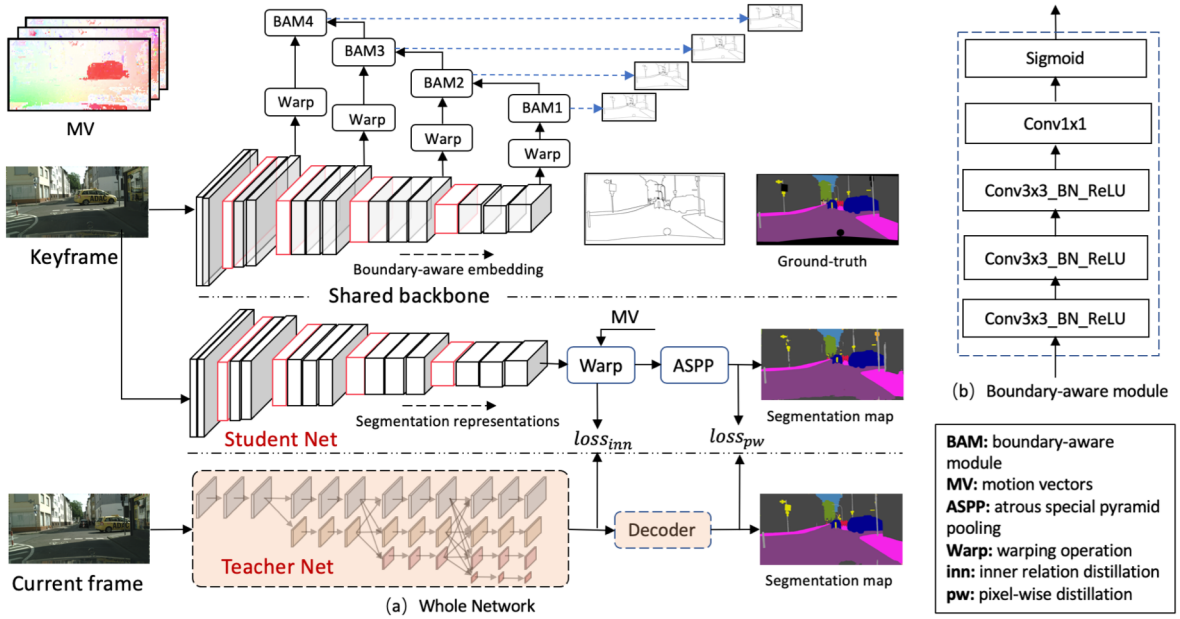


Fig. 3. The illustration of our BDNet for video semantic segmentation. (a) The top part is the boundary-aware stream. The middle one is the main stream, also known as the student network that shares the backbone with the boundary-aware stream. The bottom one is the well-trained teacher network. (b) The structure of boundary-aware module.

——Main stream

The main stream is divided into three modules: N_{feat}, N_{task}, N_{out}.

- N_{feat} takes as input a keyframe $I_{t\to\tau}$ to extract dense semantic features. With the motion vectors, warping operation is applied to propagate $f_{t\to\tau}$ to current frame I_t .
- N_{task} utilizes atrous spatial pyramid pooling (ASPP) module as decoder and returns score maps.
- N_{out} upsamples score maps to original resolution to output final segmentation results.

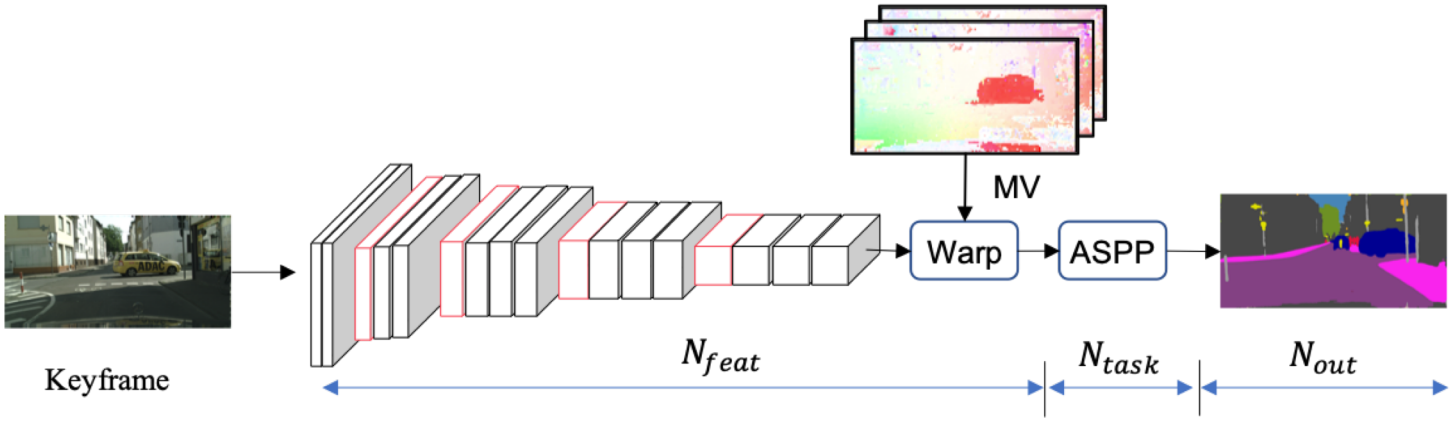
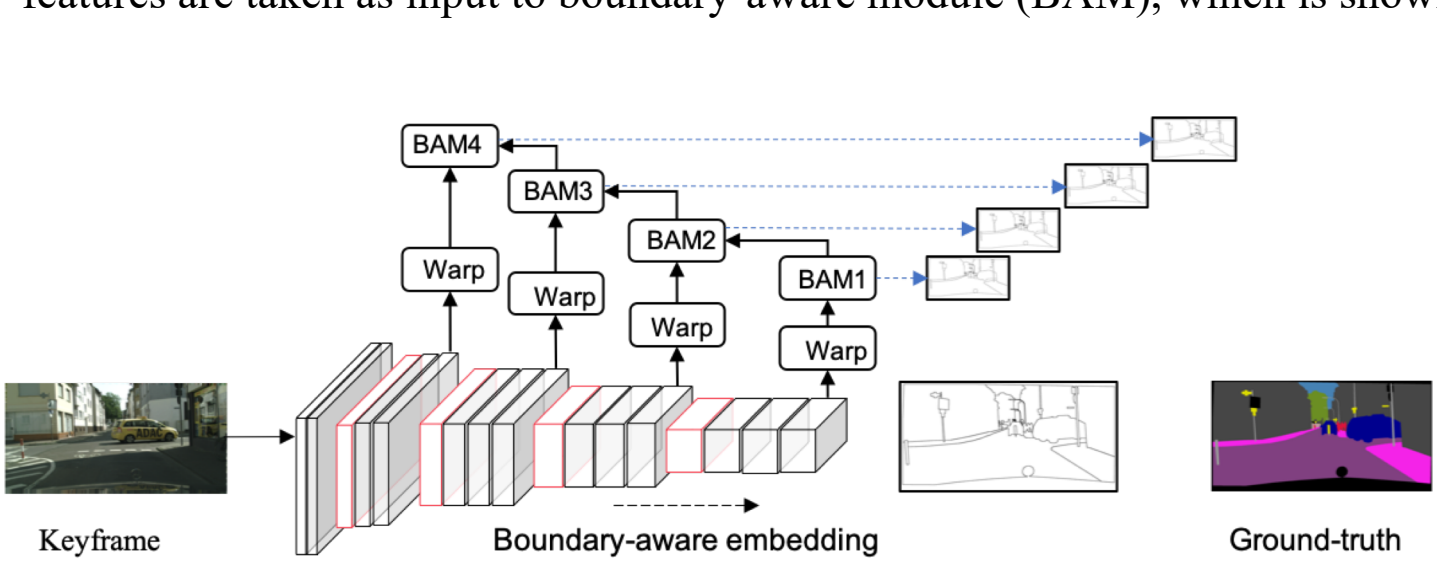


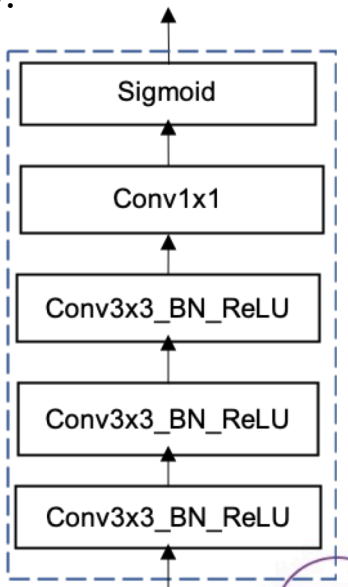
Fig. 4. The illustration of the main stream.

——Boundary-aware stream

The boundary-aware stream adopts an encoder-decoder framework, sharing the backbone with the main stream. The decoder employs top-down architecture with lateral connection. The intermediate features are warped, followed by a concatenation operation with upsampled features from top-level. The concatenated features are taken as input to boundary-aware module (BAM), which is shown in Fig.5.



(a) Boundary-aware stream



(b) Boundary-aware module

Fig. 5 The illustration of boundary-aware stream.

——Distillation stream

The teacher network takes as input the current frame I_t and returns high-quality score maps s_t^T . The student network, which is the main stream, propagates the keyframe features $f_{T\to\tau}$ forward, outputting score maps at current time t. The inner distillation is proposed to guide the main stream to learn hidden relationships in intermediate features.

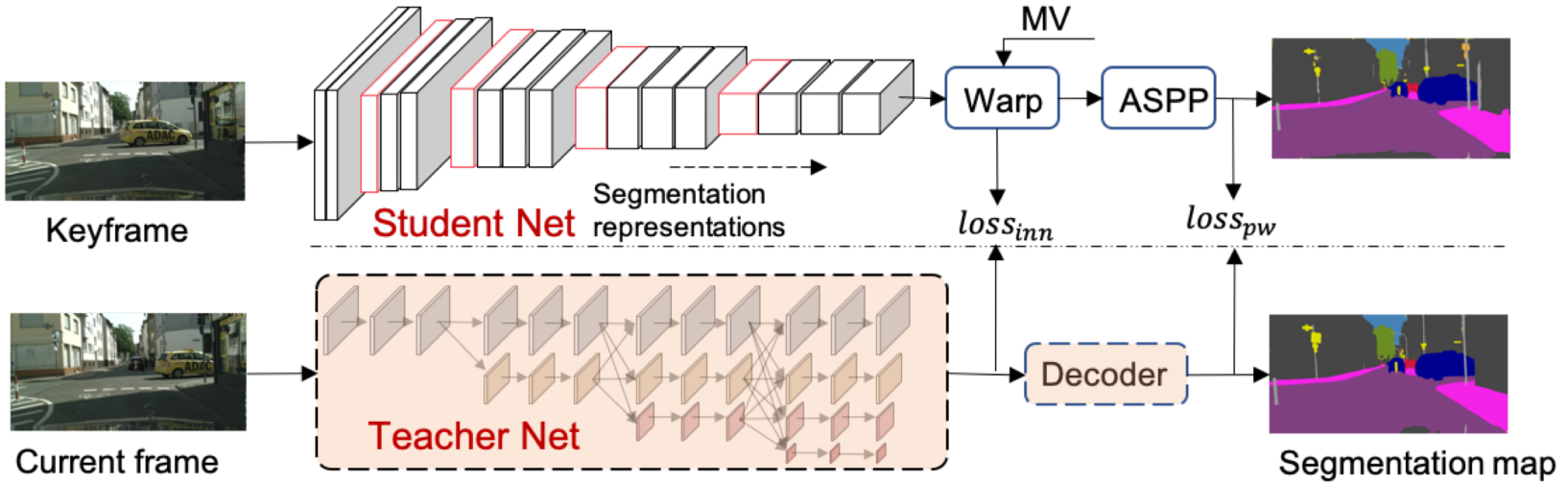


Fig. 6 The illustration of distillation stream.

The boundary-aware stream training loss

In boundary-aware stream, the training loss function is defined as the sum of three types of loss functions:

$$L_B = \sum_{k=1}^{K} l^k = \sum_{k=1}^{K} l_{BCE}^k + l_{IOU}^k + l_{SSIM}^k$$

where $l_{BCE}$is the BCE loss, l_{IOU} indicates IOU loss and l_{SSIM} is the SSIM loss.

$$l_{BCE} = -\sum_{I,J} Y_{i,j} \cdot \log\left(x_{i,j}\right) + (1 - y_{i,j}) \cdot \log(1 - x_{i,j})$$

Where $y_{i,j}$ is the ground truth at position (i, j) and$x_{i,j}$ is the predicted probability.

$$l_{IOU} = 1 - \frac{\sum_{i=1}^{h}\sum_{j=1}^{w} y_{i,j} \cdot x_{i,j}}{\sum_{i=1}^{h}\sum_{j=1}^{w}[y_{i,j} + x_{i,j} - y_{i,j} \cdot x_{i,j}]}$$

$$l_{SSIM} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\delta_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\delta_x^2 + \delta_y^2 + C_2)}$$

where $x = \{x_i : i = 1, \dots, N_2\}$ and $y = \{y_i : i = 1, \dots, N_2\}$ are pixel values of two corresponding patches $(N \times N)$ cropped from the predicted probability map and ground truth mask, μ_x, μ_y and δ_x, δ_y are mean values and standard deviation of x and y, δ is covariance and C ,C are set *to 0.012, 0.032,* respectively.

Pixel-wise and inner-relation distillation

Knowledge distillation is used to transfer knowledge from the teacher network T to the student network S. we apply pixel-wise and inner-relation knowledge distillation in video semantic segmentation task to speed up the inference with comparable accuracy.

Pixel-wise distillation. Kullback-Leibler divergence to align distributions of probability maps between main stream and teacher network.

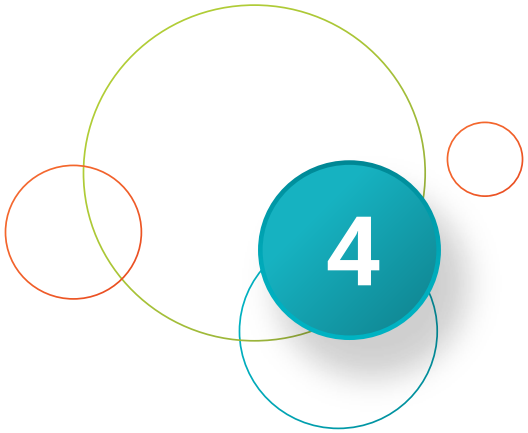
$$l_{pw} = \frac{1}{N}\sum_{i \in N} KL(p_i^S || p_i^T)$$

Inner-relation distillation. The relationship map of teacher network using non-local operation and transfer it to the student network. The loss function is defined as :

$$l_{inn} = \frac{1}{(h \times w)^2}\sum_i \sum_j \left\| \varphi_{(i,j)}^T - \varphi_{(i,j)}^T \right\|_2 \qquad \varphi_{(i,j)} = \frac{v_i}{||v_i||_2} \cdot \frac{v_j}{||v_j||_2}$$

The total loss is the sum of three losses:

$$L_{loss} = L_S + \lambda L_B + \beta L_D$$



Experimental Results

——Experiments

Datasets:

The model is trained and evaluated on video semantic segmentation dataset Cityscapes. The dataset is consisted of 30-frame snippets that collected at 17 fps with the resolution of 1024 × 2048. Each snippet is sparsely annotated at 20th frame. There are totally 30 classes with only 19 classes used for evaluation.

Evaluation metrics:

Mean intersection over union (mIoU) is used as the segmentation accuracy. To assess efficiency, we randomly selected keyframe at interval 5 and calculate speed (second per frame) over the validation dataset.

Implementation details:

We utilize DeepLabv2 as the single-frame segmentation network. The teacher network is HRNet. The input frames are randomly selected in range [τ, 1], where τ = 5 is the interval before current frame and τ = 1 means segment every frame in snippet.
At the inference stage, we test the main stream on Tesla K80 GPU, ignoring the boundary-aware stream and the teacher network.

——Experiments

We extract motion vectors from compressed videos and replace optical flow network to reduce computational cost. DFF is reimplemented on compressed videos and test performance at interval 5 on Cityscapes.

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 stream 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

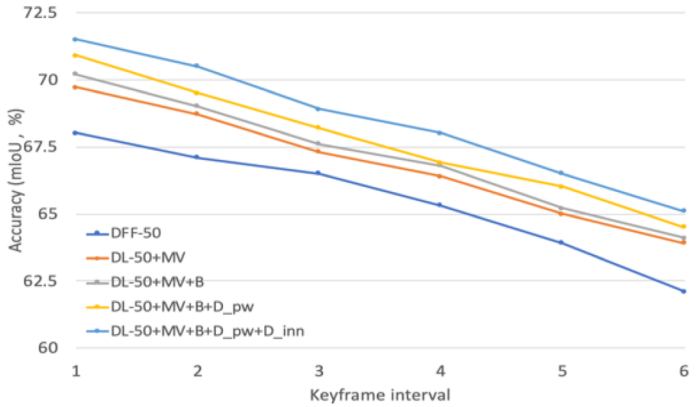


Fig. 3. Accuracy on various keyframe intervals for Cityscapes dataset. The model is based on ResNet-50.

——Results

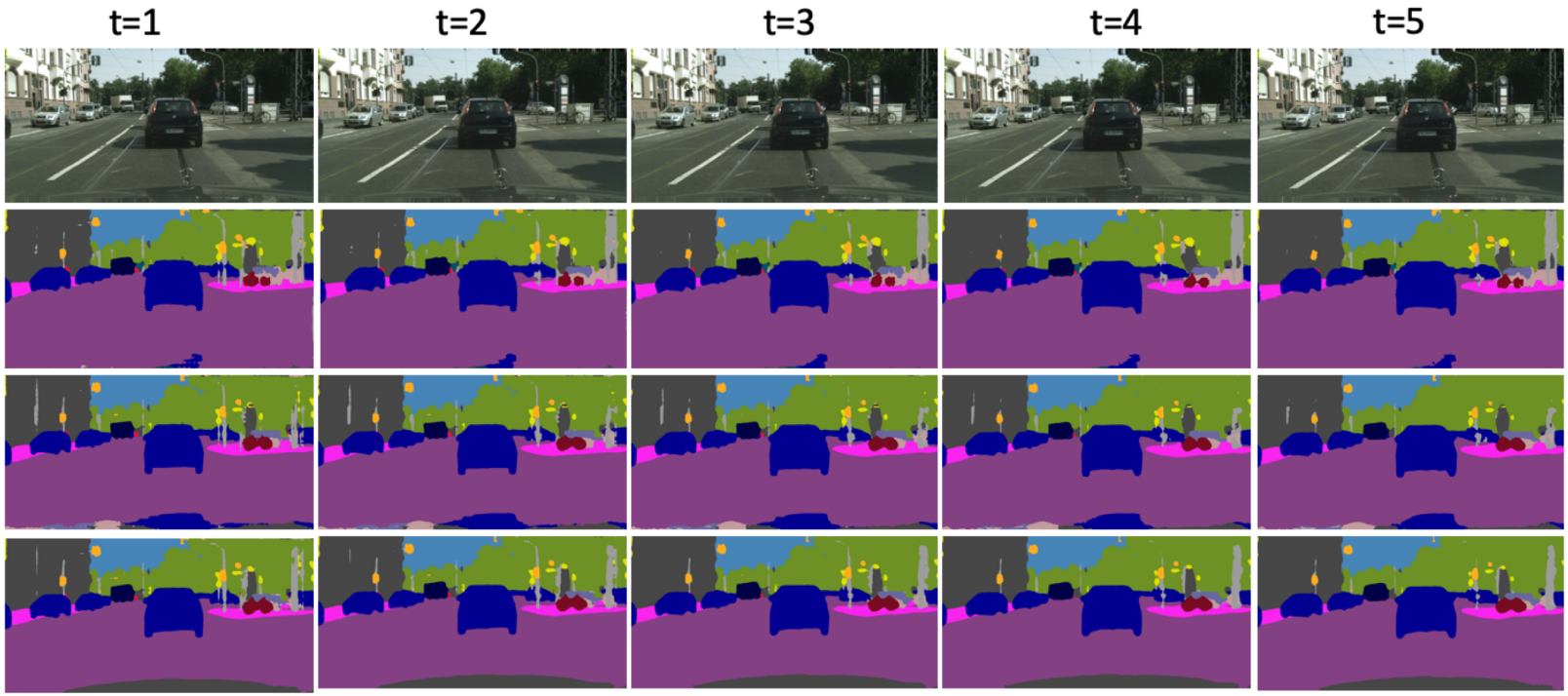


Fig 7. Visualization of video semantic segmentation results. The column denotes various keyframe intervals. The first row is input I-frame. The second one is results of DFF based on DL-50. The third one is results of DL- 50+MV+BAS. The last row shows the results of BDNet-50.

——Results

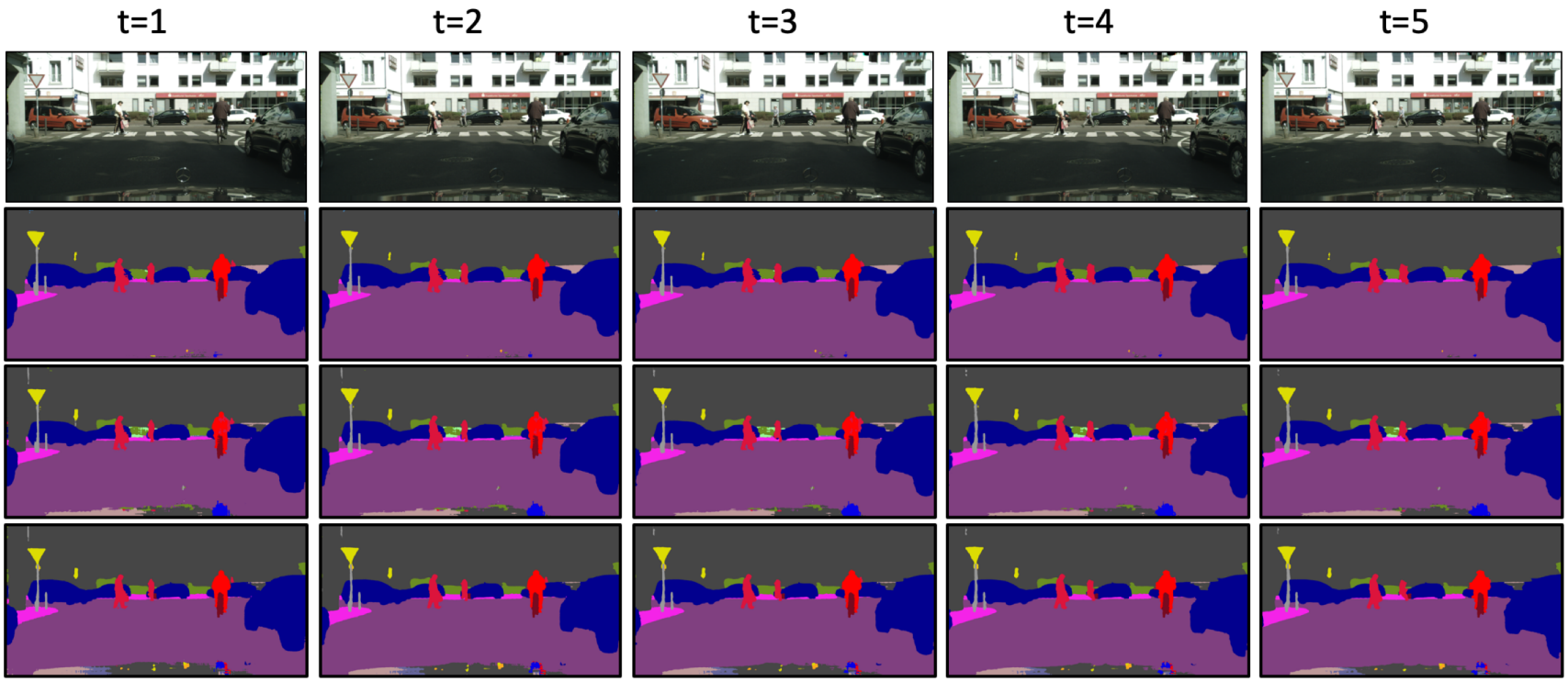
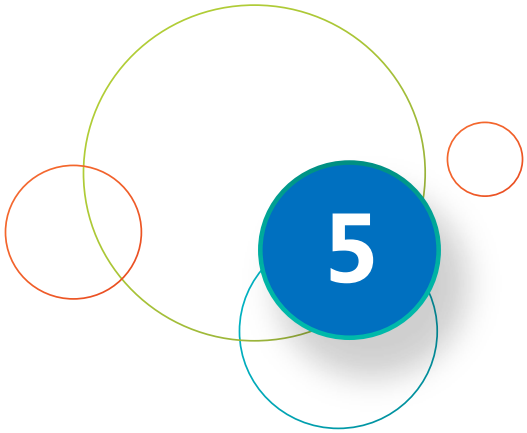


Fig 8. Visualization of video semantic segmentation results. The column denotes various keyframe intervals. The first row is input I-frame. The second one is results of DFF based on DL-50. The third one is results of DL- 50+MV+BAS. The last row shows the results of BDNet-50.



Conclusion

Conclusion:

1. The main stream.

   The main stream uses motion vectors instead of optical flow to improve efficiency.

2. The boundary-aware stream.

   The boundary- aware stream is added to predict silhouette of objects for enhancing edge features of segmented regions.

3. The distillation stream.

   A teacher network is further applied to transfer knowledge to student network, in order to correct the tailing effect caused by warping errors.

——Reference

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in CVPR, 2015, pp. 3431–3440.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. InCVPR, 2017.

[3] 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. arXiv, 2016.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv, 2017.

[5] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in CVPR, 2018, pp. 1857–1866.

[6] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, 2017, pp. 2349–2358.

[8] Y.-S.Xu,T.-J.Fu,H.-K.Yang,andC.-Y.Lee,"Dynamic video segmentation network," in *CVPR*, 2018, pp. 6556–6565.

[9] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *CVPR*, 2017, pp. 3000–3009.

[10] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation," in *CVPR*, 2017, pp. 5696–5704.

[12] S. Wang, H. Lu, and Z. Deng, "Fast object detection in compressed video," in *ICCV*, 2019.

[13] S. Jain and J. E. Gonzalez, "Inter-bmv: Interpolation with block motion vectors for fast semantic segmentation on video," *arXiv preprint arXiv:1810.04047*, 2018.

清華大學 Tsinghua University

Please contact us:

Hongchao Lu: luhc15@mails.tsinghua.edu.cn

Zhidong Deng: michael@tsinghua.edu.cn