Supplementary Materials for "Perceptual Consistency in Video Segmentation"

Yizhe Zhang [*†] yizhe.zhang.cs@gmail.com
Shubhankar Borse [‡§] sborse@qti.qualcomm.com
Hong Cai [‡§] hongcai@qti.qualcomm.com
Ying Wang [*] yingwang0022@gmail.com
Ning Bi [¶] nbi@qti.qualcomm.com
Xiaoyun Jiang [¶] xjiang@qti.qualcomm.com
Fatih Porikli [‡] fporikli@qti.qualcomm.com

1. Predicting Segmentation Correctness across Datasets

To further evaluate the efficacy of using our proposed perceptual consistency in predicting pixel-wise segmentation correctness, we conduct experiments with networks (HRNet-w18 and HRNet-w48) that are trained on Cityscapes and then tested on MIT DriveSeg. The network weights are obtained directly from the HRNet official repository.[1]

Fig. 1 and 2 show the ROC curves for the cases of HRNet-w18 and HRNet-w48, respectively, with different sparsity levels of the available ground-truth segmentation maps. It can be seen that in all cases, by combining confidence and our perceptual consistency (green), we are able to improve the prediction accuracy of the segmentation correctness, as compared to using confidence alone (blue) and combining confidence and optical flow (purple).

Fig. 3 and 4 show the precision-recall curves. It can be seen that for the cases of both networks and for different sparsity levels of the ground truths, our proposed combination of confidence and perceptual consistency (green) provides significantly better prediction performance (in terms of AUC), as compared to using confidence alone (blue) and fusing confidence and optical flow (purple). Note that a random classifier for incorrect/correct segmentation will have precision-recall AUCs of 0.162 and 0.083 for the cases of HRNet-w18 and HRNet-w48, respectively.

2. Visualizing Segmentation Error Prediction

In Fig. 5, we visually compare the true segmentation errors (first column), the confidence-based predicted error map (second column), as well as the predicted error map based on our perceptual consistency (third column). It can be seen that the confidence-based prediction and the perceptual-consistency-based prediction complement each other. This is because the confidence captures the model-internal image-level information while perceptual consistency captures the model-external temporal information.

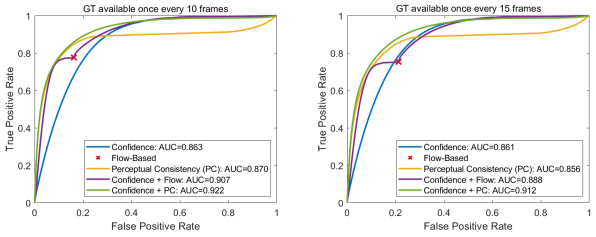

Figure 1: ROC curves for predicting pixel-wise segmentation correctness on DriveSeg, for the cases where a ground-truth segmentation map is available every 10 frames (left) and every 15 frames (right). An incorrectly segmented pixel is considered as a positive sample and a correct one is a negative sample. The segmentation model is an HRNet-w18 trained on Cityscapes.

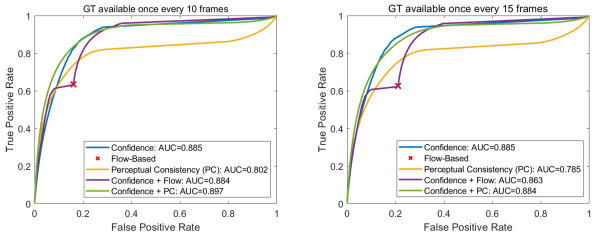

Figure 2: ROC curves for predicting pixel-wise segmentation correctness on DriveSeg, for the cases where a ground-truth segmentation map is available every 10 frames (left) and every 15 frames (right). An incorrectly segmented pixel is considered as a positive sample and a correct one is a negative sample. The segmentation model is an HRNet-w48 trained on Cityscapes.

[*] Work done at Qualcomm AI Research.
[†] Nanjing University of Science and Technology, Nanjing, China
[‡] Qualcomm AI Research, an initiative of Qualcomm Technologies, Inc.
[§] Equal contribution.
[¶] Qualcomm Technologies, Inc.
[1] The repository can be found at https://github.com/HRNet/HRNet-Semantic-Segmentation/tree/pytorch-v1.1.

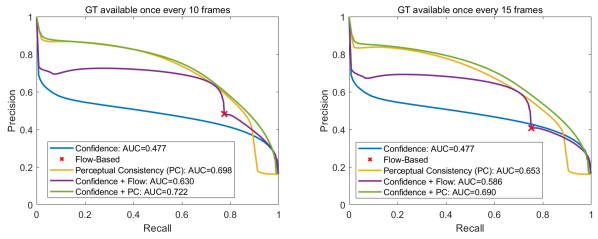

Figure 3: Precision-recall curves for predicting pixel-wise segmentation correctness on DriveSeg, for the cases where a ground-truth segmentation map is available every 10 frames (left) and every 15 frames (right). An incorrectly segmented pixel is considered as a positive sample and a correct one is a negative sample. The segmentation model is an HRNet-w18 trained on Cityscapes. The AUC of a random classifier is 0.162.

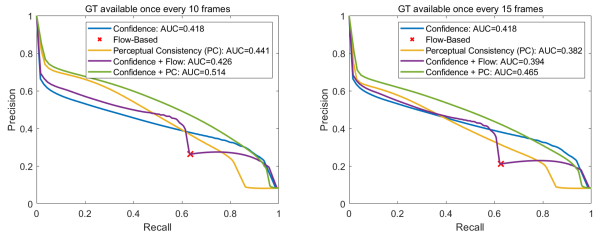

Figure 4: Precision-recall curves for predicting pixel-wise segmentation correctness on DriveSeg, for the cases where a ground-truth segmentation map is available every 10 frames (left) and every 15 frames (right). An incorrectly segmented pixel is considered as a positive sample and a correct one is a negative sample. The segmentation model is an HRNet-w48 trained on Cityscapes. The AUC of a random classifier is 0.083.

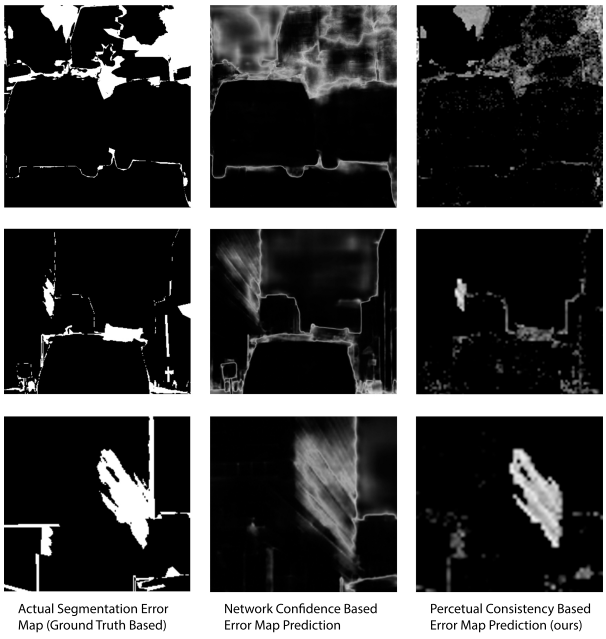

Figure 5: Visualization of ground-truth segmentation error map, confidence-based error prediction, and perceptual-consistency-based prediction. These results are based on the experiments of Sec. 1 in this supplementary file (with the HRNet-w18 model).