NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

CCDS24016 - Object Tracking-Guided Mask Correction for Consistent Real-Time Semantic Segmentation

Presented by Gan Qing Rong (U2321908E)

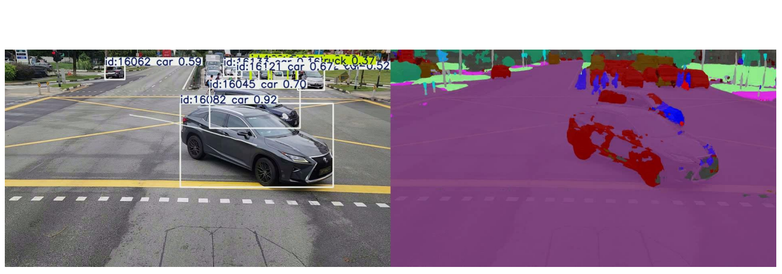
Supervised by Dr Loke Yuan Ren

Motivation

Problem:
Standard semantic segmentation models can struggle with consistency across frames in video streams. Objects may have inaccurate masks due to occlusion, motion blur, or unfavorable lighting conditions

Why It Matters:
- Poor segmentation affects downstream tasks like autonomous driving, surveillance, and augmented reality.
- Ensuring temporal consistency in real-time segmentation is crucial for real-world applications.

Limitations of Current Approaches:
- Traditional segmentation treats each frame independently, leading to inconsistencies across frames.
- Post-processing techniques like CRF or optical flow struggle with fast-moving objects.



Poor segmentation mask of a car

Methodology

We enhance semantic segmentation by integrating **YOLO11 object tracking** to improve temporal consistency:

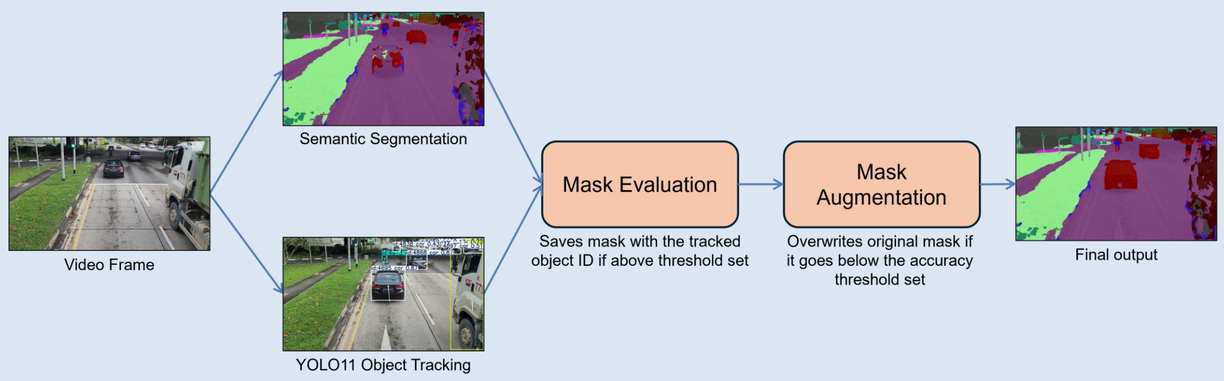
1. Segmentation & Tracking: A segmentation model generates masks, while a YOLO11 object tracker assigns object IDs and object classes.

2. Mask Evaluation: For each tracked object, we compute the proportion of correctly classified pixels within its bounding box.

3. Consistency Check: If an object has appeared before, we compare its mask consistency with past frames by measuring *change in percentage* of masked pixels which are of the same class as the object detected.

4. Mask Correction: If the consistency drops below *threshold T*, we restore the previous mask; otherwise, the current mask is stored for future use.

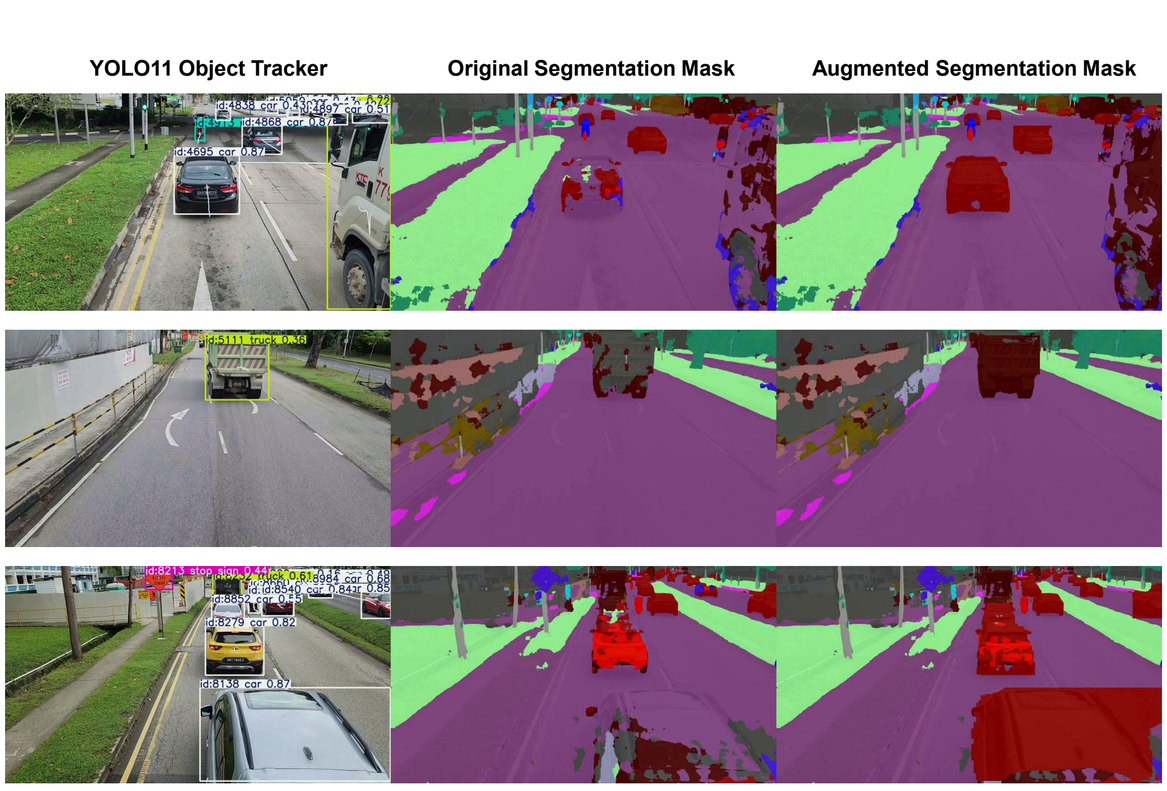
This method reduces segmentation errors caused by occlusion, motion blur, and misclassification.



Pipeline of Late-Fusion Model

Qualitative Results

We tested the DeepLabV3Plus model pretrained on Cityscapes dataset and compared its original results with that of the late fusion model. The late fusion model boasts higher accuracy and temporal consistency throughout the video, but suffers from false positives due to the slight inaccuracies of the saved masks:



Quantitative Results

We also tested the models on the VSPW dataset, which involved multiple videos downloaded from YouTube. Since the semantic segmentation models were pretrained on Cityscapes dataset, we filtered out the videos that did not contain vehicles and people. The final dataset contained 155 videos with 45 to 140 frames per video.

Below shows the quantitative results on the unseen VSPW dataset:

Model	Threshold	mIoU	Accuracy	Precision	Recall	F1 Score
MobileNetV3	NA	44.441%	51.684%	23.485%	17.445%	17.234%
MobilNetV3 + YOLO11 ObjTracker (Ours)	**0**	**45.715%**	**53.228%**	**24.111%**	**18.407%**	**17.994%**
DeepLabV3Plus	NA	46.397%	46.153%	30.912%	21.500%	21.035%
DeepLabV3Plus + YOLO11 ObjTracker (Ours)	**0**	**47.713%**	**47.466%**	**31.431%**	**22.513%**	**21.867%**
DeepLabV3Plus + YOLO11 ObjTracker (Ours)	**-5**	**47.756%**	**46.651%**	**32.015%**	**22.774%**	**22.191%**