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

Gan Qing Rong
College of Computing and Data Science

Dr. Loke Yuan Ren
College of Computing and Data Science

Abstract – Real-time semantic segmentation is challenging due to conditions such as occlusion, lighting variations, motion blur, and inconsistencies across consecutive frames. This issues are even more prevalent when using lightweight semantic segmentation models. In this paper, we propose a novel late-fusion pipeline that significantly improves the temporal consistency and accuracy of prediction masks by integrating object tracking and semantic segmentation. We utilize YOLOv11 [1] for object tracking combined with MobileNetV3 [2] and DeepLabv3+ [3] for semantic segemntation, focusing on urban environments and street scenes, such as Cityscapes. Our **Tracking-Oriented Semantic Segmentation (TOSS)** pipeline tracks each object's mask and augments new ones to mitigate temporal inconsistencies of each object's masks across consecutive frames.

Our experiments demonstrate that TOSS outperforms baseline methods such as using a sole semantic segmentation model, for example, MobileNetV3 and DeepLabV3+ across most, if not all metrics. We also identified that the method tends to have higher performance for specific classes such as persons and cars compared to trucks, buses and motorcycles.

By introducing of object tracking and semantic segmentation fusion model, we enhance the ability of semantic segmentation models and achieve higher temporal consistencies

Our pipeline is available at:

https://github.com/CobaltConcrete/mmsegmentation.git

Keywords – Semantic segmentation, object tracking, late-fusion model

1 INTRODUCTION

Semantic segmentation is a computer vision technique that assigns a class label to every pixel in an image, allowing for scene understanding. In contrast, object detection focuses on identifying objects in an image with bounding boxes or masks. Object tracking associates these detections across sequential frames to maintain and track each object's identity over time.

While powerful individually, semantic segmentation models often lack temporal awareness, leading to inconsistent predictions across frames in video streams — especially under challenging conditions such as occlusion or lighting changes [4].

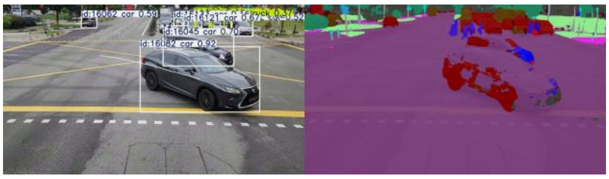

Figure 1: YOLO11 tracking (left) and an example of poor segmentation by DeepLabV3Plus (right).

To address this, we propose **Tracking-Oriented Semantic Segmentation (TOSS)**, a pipeline and method that enhances real-time semantic segmentations by combining the strengths of semantic segmentation and object tracking. By leveraging the keeping track of each unique object's masks with the YOLO11 Object Tracker's object ID, we prepare a safe previous mask to augment the next mask if it were to lose consistency across the frames.

This paper focuses specifically on semantic segmentation within urban and general real-world environments. It operates under the assumption that object tracking can reliably link instances across frames, allowing temporal information to aid the segmentation process. However, certain limitations remain: occlusions can still degrade the quality of object tracks, and false positives from prior masks may introduce errors that propagate over time.

The following sections review related work, describe our approach, and present experimental results that demonstrate the benefits of incorporating temporal information into semantic segmentation.

2 RELATED WORKS

Stanczyk, T., & Brémond, F. [5] combine bounding box and mask information for consistent multi-object tracking (MOT), using temporally propagated masks to improve tracklet-detection associations. Our work instead focuses on real-time segmentation consistency rather than MOT, using YOLO11 tracking to directly correct segmentation masks and ensure temporal coherence.

Shuo, H. et al. propose recovering missing semantic masks using temporal and depth data in SLAM systems. Nilsson, D. & Sminchisescu, C. [6] introduce a deep video segmentation method that uses FlowNet2-based optical flow and a gated recurrent unit to propagate and refine semantic labels across frames, improving both accuracy and temporal consistency over per-frame baselines. **TOSS** eliminates the need for depth sensors or computationally intensive optical flow estimation (FlowNet2) to propagate labels by relying on object tracking to resolve occlusion and motion blur and hence it is more generalizable to standard video streams.

3 METHOD

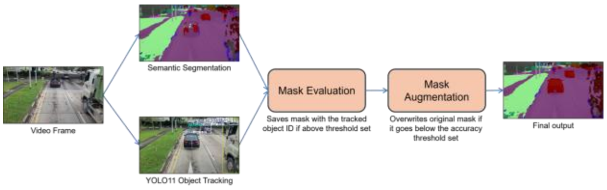
Video Frame
Semantic Segmentation
YOLO11 Object Tracking
Mask Evaluation
Saves mask with the tracked object ID if above threshold set
Mask Augmentation
Overwrites original mask if it goes below the accuracy threshold set
Final output

*Figure 2: Pipeline of **TOSS**.*

3.1 TOSS PIPELINE

TOSS is a late-fusion model consisting of a semantic segmentation model and an object tracking model. For this paper, we have chosen YOLO11 as our object tracker model due to the simplicity of its integration. We enhance the temporal consistency of real-time semantic segmentation of different models by integrating YOLO11's object tracking.

3.1.1 MESAURING PIXEL CLASSIFICATION ACCURACY WITHING OBJECT BOUNDARY

When **TOSS** receives an image frame from a video, the image is passed through both the semantic segmentation model (S) and object tracking model (O).

For each tracked object o in frame t, we compute the classification accuracy A_t^o of the segmentation mask M_t within the bounding box B_t^o:

$$A_t^o = \frac{\sum_{(x,y)\in B_t^o} 1[M_t(x,y) = C_t^o]}{|B_t^o|} \quad (1)$$

- $M_t(x,y)$ is the class at pixel (x,y) predicted by the semantic segmentation model at frame t
- C_t^o is the class assigned to object o by the object tracker model at frame t
- 1[...] is the indicator function of whether the above 2 predicted classes match
- $|B_t^o|$ is the number of pixels in the bounding box B_t^o

3.1.2 TEMPORAL MASK CONSISTENCY

To assess consistency with previous frame $t-1$, we define the class-consistent pixel change ratio:

$$\Delta A_t^o = \frac{A_t^o - A_{t-1}^o}{A_{t-1}^o} \quad (2)$$

3.1.3 MASK CORRECTION

We restore the object's mask M_{t-1}^o from the previous frame if the object's previous mask M_{t-1}^o **exists**, **and** consistency ΔA_t^o drops below a threshold T. Otherwise, we keep the object's new mask M_t^o. The overwriting mask is warped and resized to fit the new bounding box and the same object as closely as possible.

$$M_t^o = \begin{cases} M_{t-1}^o, & if\ \Delta A_t^o < T \\ M_t^o, & \text{otherwise} \end{cases} \quad (3)$$

3.1.4 MASK TRACKER

If a new mask was accepted for object o, we then store its new mask M_t^o and new classification accuracy A_t^o tagged to its object o.

3.2 MODELS

3.2.1 SEMANTIC SEGMENTATION MODEL

For semantic segmentation, we have chosen 2 kinds: DeeplabV3 Plus because it boasts very high performance compared to other models, as well as MobileNetV3 for a lightweight example.

These 2 models have been pretrained on the Cityscapes [7] dataset with the following classes:

Table 1: Cityscapes dataset classes.

GROUP	CLASSES
FLAT	road, sidewalk, parking, rail track
HUMAN	person, rider
VEHICLE	car, truck, bus, caravan, trailer, train, motorcycle, bicycle
CONSTRUCTION	building, wall, fence, guard rail, bridge, tunnel
OBJECT	pole, pole group, traffic light, traffic sign
NATURE	vegetation, terrain
SKY	sky
VOID	unlabelled, ego vehicle, rectification border, out or roi, static, dynamic, ground

3.2.2 OBJECT TRACKING MODEL

We use YOLO11 for object tracking due to its ease of integration and high performance despite being lightweight, making it suitable for fusion in this pipeline. YOLO11 has been pretrained on COCO dataset with up to 80 different classes.

3.2.3 SHARED CLASSES BETWEEN MODELS

Since both the semantic segmentation model and the object tracking model are trained on different classes, we defined the ***shared_classes*** between a Cityscapes-pretrained semantic segmentation model and a COCO-pretrained object tracking model: **person, bicycle, car, motorcycle, truck, bus, vegetation**.

3.3 DATASET

We decided to obtain 2 different types of datasets: one collected from real-world settings and another sourced from an established benchmarking dataset.

3.3.1 REAL-LIFE VIDEOS

Real-world data was collected from top-deck double-decker bus rides across different times of the day and varying weather conditions.

Videos were captured during sunny afternoons and rainy nights along the route between Nanyang Technological University and Jurong Point in Singapore. Furthermore, late-evening trips from Kent Ridge to Orchard Road were recorded to assess performance in settings with long shadows caused by low-angle lighting.

The dataset covers a range of lighting and environmental conditions—including day, evening, and nighttime scenes, in both sunny and rainy settings—to provide diverse and challenging scenarios for model evaluation.

3.3.2 VSPW DATASET

We utilized the **VSPW** dataset for benchmarking, which consists of videos sourced from YouTube [8]. It covers a wide range of real-world scenarios, making it a much more extensive dataset for benchmarking between lone semantic segmentation models and out **TOSS** pipeline. The dataset was filtered to include only sequences that contain both vehicles and people, aligning with the target classes used for evaluation. This filtering step leveraged the fact that the semantic segmentation model was pretrained on **Cityscapes**, focusing on the overlapping classes present in both the segmentation and tracking pipelines.

Initially, the VSPW dataset comprised a wide range of videos with varying settings and domains. To ensure compatibility between the YOLOv11 object tracker and the semantic segmentation models, we identified the common classes that can be detected by both models. This set of common classes was defined as ***shared_classes***.

The YOLOv11 object tracker was then used to detect instances of ***shared_classes*** across the dataset, allowing us to select only those videos that contain at least one of these ***shared_classes***. The final curated dataset comprised **155 videos**, each containing approximately **45–140 frames**.

The usage of this subset of VSPW dataset for benchmarking ensures consistency between the semantic segmentation and object tracking methods, making it suitable for benchmarking across varied environments.

4 EXPERIMENTS

4.1 QUALITATIVE RESULTS

We evaluated the performance of the DeepLabV3Plus model pretrained on the Cityscapes dataset and compared the results between baseline methods (using only a DeepLabV3+ model by itself) and **TOSS** pipeline. Our **TOSS** pipeline boasts better performances, being able to continue detecting and tracking persons and vehicles across frames consistently when the baseline method was unable to do so.

However, it occasionally suffered from false positives due to slight inaccuracies in the saved masks.

4.1.1 DAYTIME

During midday conditions, when lighting is optimal and shadows are minimal, the **TOSS** pipeline delivers noticeably improved qualitative results compared to using only a semantic segmentation model as seen in Figure 3. The well-lit scene allows for more accurate detection and tracking of objects, making this an ideal setting for evaluation.

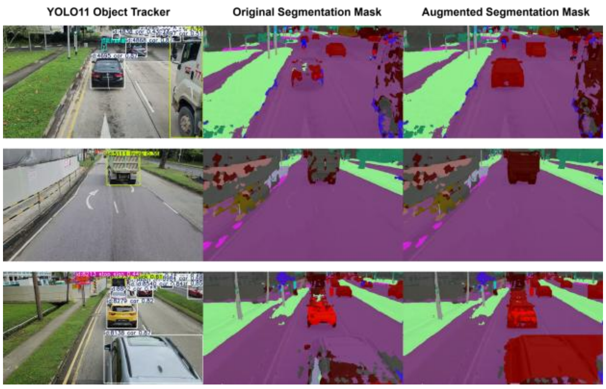

*Figure 3: Daytime comparison – Object Tracking Model (left) vs. Semantic Segmentation Model (middle) vs. **TOSS** Pipeline (right).*

4.1.2 EVENING

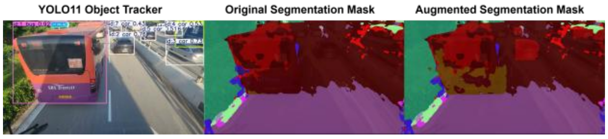

Figure 4: Late evening comparison

In late evening conditions, long shadows and higher contrast lighting introduce more challenging scenarios. The **TOSS** pipeline managed to detect the cars overtaking the bus while the lone segmentation model did not, highlighting **TOSS**' strength in low-light and challenging lighting conditions as seen in Figure 4.

4.1.3 NIGHTTIME

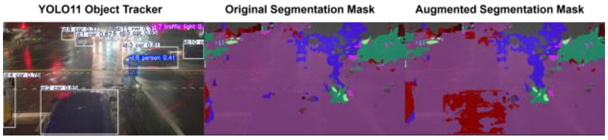

Figure 5: Nighttime comparison with rain

At night, compounded by rain and headlight glare, both the standalone semantic segmentation and the **TOSS** pipeline experience degraded performance. However, the **TOSS** pipeline still outperforms the sole semantic model, providing more accurate results and demonstrating its resilience in harsh, low-visibility environments, as seen in Figure 5, where it can still detect the nearest car as well as the two cars ahead crossing at the junction.

4.2 QUANTITATIVE RESULTS

4.2.1 GENERAL RESULTS ACROSS MODELS AND PIPELINES

We performed quantitative evaluations of both methods across a range of lighting and environmental conditions. The results, summarized in Table 1, highlight the improvements in performance of **TOSS'** late fusion approach over the baseline DeepLabV3Plus model. Metrics such as mIoU, precision, and recall were used to assess performance.

*Table 2: Performances of semantic segmentation models and **TOSS** pipeline on VSPW dataset (bolded means better performance)*

		Performance (%)		
Model	Threshold	mIoU	Acc	Prec
MobileNetV3	NA	44.4	51.7	23.5
MobileNetV3 + YOLO11 (Ours)	**0%**	**45.7**	**53.2**	**24.1**
DeepLabV3Plus	NA	46.4	46.2	30.9
DeepLabV3Plus + YOLO11 (Ours)	**0%**	**47.7**	**47.5**	**31.4**
DeepLabV3Plus + YOLO11 (Ours)	**-5%**	**47.8**	**46.7**	**32.0**

		Performance (%)	
Model	Threshold	Recall	F1 Score
MobileNetV3	NA	17.4	17.2
MobileNetV3 + YOLO11 (Ours)	**0%**	**18.4**	**18.0**
DeepLabV3Plus	NA	21.5	21.0
DeepLabV3Plus + YOLO11 (Ours)	**0%**	**22.5**	**21.9**
DeepLabV3Plus + YOLO11 (Ours)	**-5%**	**22.8**	**22.2**

The late fusion model consistently outperformed the baseline across all metrics, providing a robust solution for challenging environments with varying lighting conditions.

4.2.2 CLASS-SPECIFIC RESULTS

*Table 3: Performances of semantic segmentation models and **TOSS** pipeline on specific classes in VSPW dataset. All normal methods only include a sole DeeplabV3+ model finetuned on Cityscapes (bolded means better performance)*

Class	mIoU (%)	
	normal	**TOSS**
Person	42.4	**42.7**
Car	46.9	**49.9**
Truck	28.9	**32.0**
Bus	40.7	**56.8**
Motorcycle	23.1	**28.3**

Class	Accuracy (%)		Precision (%)	
	normal	**TOSS**	**normal**	TOSS
Person	**94.7**	94.5	**53.5**	51.5
Car	92.0	**92.4**	**73.0**	72.0
Truck	83.3	**83.5**	44.2	**44.6**
Bus	87.8	**90.5**	88.6	**88.9**
Motorcycle	94.7	**95.0**	**65.3**	61.4

Class	Recall (%)		F1 Score (%)	
	normal	**TOSS**	normal	**TOSS**
Person	67.1	**71.7**	53.8	**55.1**
Car	57.8	**63.1**	57.1	**60.2**
Truck	58.1	**62.3**	38.4	**41.4**
Bus	43.8	**62.2**	49.4	**66.3**
Motorcycle	33.4	**40.6**	31.7	**38.9**

We conducted a class-wise performance comparison for the following shared classes: **person, car, truck, bus,** and **motorcycle**, using DeepLabV3+ as the baseline. Overall, the **TOSS** pipeline improves both the **mIoU** and **F1-scores** compared to using a sole semantic segmentation model. However, a slight degradation in **precision** is observed for trucks and buses, primarily due to an increased rate of false positives as a result of their large sized masks being kept in memory.

5 CONCLUSION

This study addressed the challenges associated with real-time semantic segmentation, especially in complex urban environments and across varied lighting and weather conditions. We proposed a late-fusion pipeline **TOSS** that combines object tracking (using YOLOv11) with semantic segmentation (using MobileNetV3 and DeepLabv3+) to improve temporal consistency and prediction accuracy.

We evaluated qualitatively and quantitatively across different times of the day and weather conditions, against baseline methods of using a sole semantic segmentation model. Out **TOSS** method provides a more robust performance across almost all metrics, allowing for better real-world applications such as autonomous driving, medical imaging, and surveillance.

In summary, this work advances the state-of-the-art semantic segmentation models by introducing a pipeline that improves the quality and reliability of lightweight semantic segmentation models in challenging conditions. By leveraging object tracking to maintain a memory of detected instances across frames, we achieved more stable and accurate segmentations, addressing a key limitation in prior approaches. This contribution highlights the potential for integrating temporal information and detection-aware pipelines into future semantic segmentation models.

5.1 FUTURE WORK

5.1.1 OCCLUSION HANDLING

There is room for development of more sophisticated approaches to manage object occlusion and depth ordering of the objects in the input image or video. Incorporating depth information or scene geometry constraints can help prioritise masks when there is an overlap in objects.

5.1.2 REFINED MASK AUGMENTATION

Improvments can be made to the augmentation and overwriting of masks between frames to reduce false positives caused by misaligned masks. Techniques such as temporal consistency constraints, optical flow–guided updating, or learned mask blending may help mitigate these errors.

5.1.3 EXTENDED CALSS COVERAGE

The set of overlapping classes ***shared_classes*** between detection and segmentation pipelines can be expanded for richer scene understanding. The semantic segmentation model and object tracker may be finetuned using the same set of classes as well.

ACKNOWLEDGEMENT

Thanks to Dr. Loke Yuan Ren who gave helpful advice in conducting experiments.

Thanks to NTU's REP Makers' NVIDIA GeForce RTX 4090 which hosted 90% of my entire project.

REFERENCES

[1] Khanam, R., & Hussain, M. (2024, October 23). YOLOv11: An overview of the key architectural enhancements (arXiv:2410.17725) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.17725

[2] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019, May 14). *Searching for MobileNetV3: Efficient model architecture by combining complementary search techniques* (arXiv:1905.02244) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.02244

[3] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018, August 22). Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation (arXiv:1802.02611) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1802.02611

[4] Baghbaderani, R. K., Li, Y., Wang, S., & Qi, H. (2024, January). *Temporally-consistent video semantic segmentation with bidirectional occlusion-guided feature propagation.* In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 685–695). https://openaccess.thecvf.com/content/WACV2024/papers/Baghbaderani_Temporally-Consistent_Video_Semantic_Segmentation_With_Bidirectional_Occlusion-Guided_Feature_Propagation_WACV_2024_paper.pdf

[5] Huai, S., Cao, L., Zhou, Y., Guo, Z., & Gai, J. (2025). A multi-strategy visual SLAM system for motion blur handling in indoor dynamic environments. Sensors, 25(6), 1696. https://doi.org/10.3390/s25061696

[6] **Stanczyk, T., & Brémond, F. (2024, September 25).** *Temporally propagated masks and bounding boxes: Combining the best of both worlds for multi-object tracking*

(arXiv:2409.14220v1) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.14220

[7] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019, May 14). *Searching for MobileNetV3: Efficient model architecture by combining complementary search techniques* (arXiv:1905.02244) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.02244

[8] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016, April 7). *The Cityscapes Dataset for Semantic Urban Scene Understanding* (arXiv:1604.01685v2) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1604.01685