Waymo Panoramic Video Panoptic Segmentation Challenge Entry: Clip kMax

Venkata Pradeep Kadubandi Aniket Murarka

I. Introduction

This year, as part of the Waymo Open Dataset (WOD) set of challenges, a new challenge Panoramic Video Panoptic Segmentation (PVPS) was introduced [10]. This challenge requires video panoptic segmentation to be consistent across multiple synchronized camera streams. Particular to the dataset, there are 5 cameras mounted on the vehicle viewing the scene from the left to the right. Instance categories such as pedestrians, vehicles, and cyclists are tracked across both cameras and time.

Our approach to the challenge combines Clip-kMax that leverages k-Means cross attention across a video clip, as described in the Video kMax paper [13] with Video Panoptic Stitching [11] for post processing. While our current implementation does not include the HiLA-MB module [13], we discuss its potential as future work. Besides, we propose Panoramic Camera Stitching across multi-camera frames as future work.

This technical report is organized as follows. In section II, we review some of the literature for Panoptic Segmentation. Our experiments and results are described in section III. In the final section IV, we discuss challenges, work in progress and further directions.

II. Related Work

In recent years, there has been significant progress in panoptic segmentation, particularly with the advent of transformer attention mechanisms. We have extensively studied the following state-of-the-art methods most of which leverage transformer attention mechanisms.

- DeTR [1]: DeTR proposes a transformer-based object detection and panoptic segmentation framework, which replaces heuristics like anchor box proposals with direct box predictions by introducing object queries. It achieves impressive performance by directly predicting objects and their attributes.
- MaskFormer [5]: MaskFormer extends the vision transformer architecture for panoptic segmentation from DetR by adopting a pixel decoder. It differs from Max-DeepLab [15] in that it does not require auxiliary training losses.
- Mask2Former [7]: Mask2Former improves upon MaskFormer by using masked attention in the decoder model which restricts attention to local features. Additionally, Mask2Former is easily extendable to video panoptic segmentation by simply stacking 2 consecutive image frames [6].
- Segment Anything [9]: We explored using the Segment Anything model published very recently. This is a promptable model that allows zero-shot learning (through composition, etc) for various tasks such as edge detection and instance segmentation. Unfortunately, one of the shortcomings of the work is the lack of simple prompts for panoptic segmentation. Regardless, using the Segment Anything model as a backbone is a future direction we would like to pursue.
- DeepLab Family [2, 3, 4]: The DeepLab series of papers have made significant contributions to semantic segmentation and object detection. By incorporating dilated convolutions and atrous spatial pyramid pooling, they have achieved remarkable results in various challenging scenarios.
- Panoptic DeepLab [8]: Panoptic DeepLab extends the DeepLab framework to incorporate panoptic segmentation. The method performs instance segmentation through instance center prediction and regression branches and combines them with semantic segmentation from DeepLab family to achieve state-of-art bottom up panoptic segmentation.
- ViP DeepLab [11]: ViP DeepLab extends Panoptic DeepLab to the video domain by adding additional branches for center regression of the next frame in the video sequence w.r.t the current frame. We experimented with ViP DeepLab architecture in addition to clip kMax for solving the challenge.
- Max DeepLab [15]: Max DeepLab introduces a novel dual-path transformer architecture that leverages several possible forms of attention between pixel path and memory path. The PQ style loss and other auxiliary losses are introduced in this work.
- K-Means Mask Transformer [16]: This recent approach combines K-means clustering and transformer-based attention mechanisms for panoptic segmentation. This work made a novel connection between k-means clustering and transformer attention mechanisms.
- Video kMax [13]: This work extends K-Means Mask Transformer to the video domain by simply concatenating images in a clip along the height axis generating a tube prediction over the clip. **The simplicity of the approach and the state-of-the art performance motivated us to attempt this method for the challenge**. The work

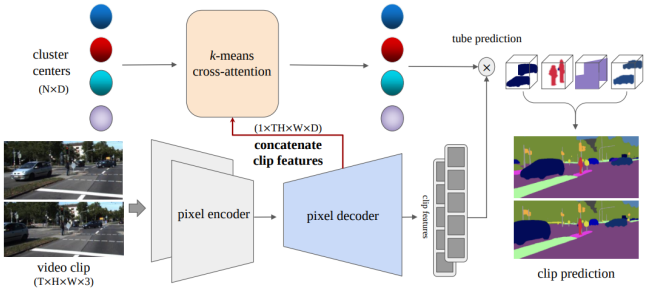

Fig. 1. Clip kMax architecture. Our implementation uses T = 2. Reference [13]

also proposes a Hierarchical Location Aware Memory Buffer (HiLA-MB) for object tracking across long video sequences. Our current implementation does not include the HiLA-MB module but we believe it's a promising direction to explore further.

III. Method & Results

Our approach, Clip kMax as introduced in [13], leverages K-Means cross attention to address the WOD 2D Video Panoptic Segmentation challenge. As mentioned in the original paper, we used a clip of length 2. The clip images are concatenated along the height axis and fed as input to the model during both training and inference.

Our model takes images from all different cameras similar to the 'View' mode training in the WOD PVPS paper [10]. We generated source data as pairs of images at consecutive timesteps. For training and validation datasets, this refers to images at consecutive time steps where the ground truth labels are available. For the test dataset, for inference, we used the image at the very next timestamp as the second image of the clip.

Due to limited GPU availability within the timelines of the competition, we conducted our experiments using a machine equipped with a single A100 GPU, which had a memory limit of 40 GB. To accommodate the memory constraints, we employed a quarter crop size of the total clip dimensions for training the Clip kMax model. For example, the original clip size for 3 cameras was 2560 x 1920, while our crop size was set to 641 x 481. During training, we utilized a mini-batch size of 4.

To implement our method, we adapted the existing open-source DeepLab implementation [2], by Google Research. Specifically we used the K-Max DeepLab architecture with ConvNext-L backbone. To feed the clip as an Image, we developed a decoder implementation that concatenates the clip frames along the height axis to provide the input image for the model. We could not use a pre-trained backbone as we faced issues with checkpoint loading that are yet to be figured out. We also experimented with a variant of ViP-DeepLab (trained with the same training compute constraints) that used a higher mini-batch size (8), crop size (half crop of image compared to quarter crop of clip) and Resnet backbone that is pre-trained on ImageNet. Our Clip kMax training used random scale, auto augmentation (implemented in DeepLab as a simple classification policy) and panoptic copy paste as augmentations. We set the min and max resize values the same as crop size following the other kMax configurations. Our ViP-DeepLab variant only used a random scale for augmentations and no resizing. We achieved better mIoU with Clip kMax model but slightly below wAQ than the ViP DeepLab experiment. Metrics from our different experiments are outlined in Table I.

We did not get to implementing the HiLA-MB module for tracking the objects across the video sequence within the time period for the challenge. Instead we leveraged the existing Video Panoptic Stitching method provided in the DeepLab repository as a post-processing step for video stitching. The Clip kMax model generates consistent Instance Ids for a given clip of (I_t, I_{t+1}) frames. We assume a frame stitcher that can propagate Instance Ids from a concatenated panoptic frame to the next panoptic frame based on IoU matching. We refer to the [11] for additional details about this stitching. Over a video sequence, we use the frame stitcher to propagate the Instance Ids from consecutive I_{t+1} frames in the video [14]. While this is not exactly the same method employed in ViP-DeepLab, exploratively we found this to yield reasonable results compared to another alternate implementations we came up with in the short time available.

Method	mIoU	wAQ	wSQ
Clip kMax + our VPS	*27.48*	2.51	*8.31*
Vip DeepLab + VPS	25.65	*2.60*	8.16

TABLE I
Results on Waymo Test set from our experiments.

IV. Future directions and discussion

Through our work on the Waymo Open Dataset Challenge 2D Video Panoptic Segmentation using the Clip kMax method, we identified several potential future improvements to our approach. We list some of these below including works in progress:

- Increased Compute Resources: Due to the limited GPU availability during our experiments, we were constrained by the memory limit and had to compromise on the crop size and mini-batch size. One of our primary future goals is to acquire more GPUs and leverage higher computational resources. This would enable us to train the Clip kMax model with larger mini-batch sizes and potentially higher crop sizes potentially improving the performance of our approach. (*Update post competition deadline*: We already started making progress in this direction and on a promising path to better results.)
- HiLA-MB for Panoptic Stitching: From our observations, one of the shortcomings is tracking over a longer period

of time. We believe HiLA-MB [13] would improve our results significantly.

- Camera Panoramic Stitching: We worked on a module to propagate Instance Ids across cameras at the same timestep. Unfortunately, due to a bug we were not able to generate results with this module incorporated (which lead to lower than expected wAQ and wSTQ metrics). The method is similar to the one used in [10]. In particular, for each pair of cameras with overlapping FOVs, we compute IoUs between all instances in the same category. The IoUs are computed in panoramic view composed of the two camera images by equi-rectangular projection. The IoUs are then used to create a similarity matrix between all instance pairs. We use a Linear Sum Assignment algorithm [12] to find the optimal pair of matching instances between the two cameras. We do this for all neighboring cameras and propagate the Instance Ids from the leftmost to the rightmost camera. For each instance, we pick the Id from the camera with a longer history of tracking the instance. The Id is then propagated in the other camera forward in time.
- Exploring Different Attention Mechanisms: In future explorations, we aim to experiment with various attention mechanisms, such as those introduced in recent literature, to investigate their effect on the accuracy and robustness of panoptic segmentation in video sequences. One such example is Masked Attention introduced in [7].

V. Acknowledgements

We would like to the thank the Waymo Open Dataset team for providing the excellent PVPS dataset and also for their very prompt and continuous support throughout the challenge timeline.

References

[1] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV].

[2] Liang-Chieh Chen et al. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. 2017. arXiv: 1606.00915 [cs.CV].

[3] Liang-Chieh Chen et al. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. 2018. arXiv: 1802.02611 [cs.CV].

[4] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: 1706.05587 [cs.CV].

[5] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. *Per-Pixel Classification is Not All You Need for Semantic Segmentation*. 2021. arXiv: 2107.06278 [cs.CV].

[6] Bowen Cheng et al. *Mask2Former for Video Instance Segmentation*. 2021. arXiv: 2112.10764 [cs.CV].

[7] Bowen Cheng et al. *Masked-attention Mask Transformer for Universal Image Segmentation*. 2022. arXiv: 2112.01527 [cs.CV].

[8] Bowen Cheng et al. *Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation*. 2020. arXiv: 1911.10194 [cs.CV].

[9] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV].

[10] Jieru Mei et al. *Waymo Open Dataset: Panoramic Video Panoptic Segmentation*. 2022. arXiv: 2206.07704 [cs.CV].

[11] Siyuan Qiao et al. *ViP-DeepLab: Learning Visual Perception with Depth-aware Video Panoptic Segmentation*. 2020. arXiv: 2012.05258 [cs.CV].

[12] SciPy. *Linear Sum Assignment*. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.linear_sum_assignment.html (visited on 05/31/2023).

[13] Inkyu Shin et al. *Video-kMaX: A Simple Unified Approach for Online and Near-Online Video Panoptic Segmentation*. 2023. arXiv: 2304.04694 [cs.CV].

[14] TeamAniketAndPradeep. *Video Stitching for Clip kMax setup*. URL: https://github.com/PradeepKadubandi/deeplab2/blob/pk/fixes/video_inference_script_kmax.py (visited on 05/31/2023).

[15] Huiyu Wang et al. *MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers*. 2021. arXiv: 2012.00759 [cs.CV].

[16] Qihang Yu et al. *K-Means Mask Transformer*. 2023. arXiv: 2207.04044 [cs.CV].