Supplementary Materials for TubeFormer-DeepLab: Video Mask Transformer

Dahun Kim[1,3*] Jun Xie[3] Huiyu Wang[2] Siyuan Qiao[3] Qihang Yu[2] Hong-Seok Kim[3]
Hartwig Adam[3] In So Kweon[1] Liang-Chieh Chen[3]
[1]KAIST [2]Johns Hopkins University [3]Google Research

In the supplementary materials, we provide more detailed experimental results, along with more visualizations (and also video prediction results) for several video segmentation datasets. We also discuss the current limitations in the proposed TubeFormer-DeepLab in the hope to inspire more future research on the unification of video segmentation tasks. We use 'TF-DL' to denote TuebeFormer-DeepLab in the results.

1. More Experimental Results

In this section, we provide more experimental results, comparing our methods with *published* works in detail. We do not include the *unpublished and concurrent* ICCV 2021 challenge entries, which usually adopt complicated pipelines, *e.g.*, model ensembles, separate models for different sub-tasks (*e.g.*, tracking, and segmentation), multi-scale inference, or pseudo labels. In the tables, we explicitly list the adopted backbones and decoders for a detailed comparison. We note that most of the state-of-the-art approaches for different video segmentation tasks have fundamentally diverged, while our proposed TubeFormer-DeepLab is a simple and unified system for general video segmentation tasks.

[VPS] Tab. 1 summarizes our results on KITTI-STEP *val* set. As shown in the table, our TubeFormer-DeepLab-B1, employing ResNet-50 [12] and axial-attention [22], significantly outperforms Motion-DeepLab [24] (w/ ResNet-50, dual-ASPP [6] and dual decoders [7]) and VPSNet [14] (w/ ResNet-50, FPN [16], and Mask R-CNN [11] multi-head predictions) by **+12** and **+14** STQ, respectively. We also report the results in the VPQ metric [14] (another popular video panoptic segmentation metric). Similarly, our model performs better than Motion-DeepLab and VPSNet by **+11.1** and **+8.1** VPQ.

[VSS] In Tab. 2, we report our results on VSPW *val* set. As shown in the table, our TubeFormer-DeepLab-B1, employing ResNet-50 and axial-attention, significantly outperforms TCB [18] (w/ spatial-temporal OCRNet [28] and a novel memory scheme) by **+21.1** mIoU. Our TubeFormer-DeepLab-B1 also shows better results in terms of VC8 and VC16 (another video semantic segmentation metrics proposed in [18]).

[VIS] Tab. 3 summarizes our results on Youtube-VIS-2019 *val* set, along with several state-of-the-art methods.

Among the methods that predict non-overlapping segmentation, our TubeFormer-DeepLab-B1 (per-pixel), employing ResNet-50 and axial-attention, outperforms STEm-Seg [1] (using ResNet-50, FPN, and their novel 3D convolution-based TSE decoder with multi-head predictions) by **+5.8** AP. Our TubeFormer-DeepLab-B1 (per-pixel) is also better than STEm-Seg with ResNet-101 backbone by **+1.8** AP. If we also increase our backbone capacity, our TubeFormer-DeepLab-B4 (per-pixel) performs better than STEm-Seg w/ ResNet-101 by **+10.8** AP.

Our TubeFormer-DeepLab-B1 (per-pixel) performs worse than other state-of-the-art methods, including MaskProp [2], Seq Mask R-CNN [15], and the concurrent work IFC [13], since our per-pixel inference scheme generates non-overlapping predictions (*i.e.*, only one prediction for each pixel in the final output), which is disfavored by the track AP metric. To bridge the gap, we adopt the mask-wise merging scheme (denoted as per-mask) [8, 27], where each object query generates a mask proposal. The per-mask scheme significantly improves over the per-pixel scheme by more than 2 AP in the TubeFormer-DeepLab framework. Our large model TubeFormer-DeepLab-B4 with per-mask scheme outperforms MaskProp, VisTR, and IFC, and performs comparably with the best model Seq Mask R-CNN, which relies on STM [19]-like structure to propagate mask proposals through the whole sequence.

Notably, our model yields the best AR_1 and AR_{10} (**+3.9** and **+3.0** AR better than the second best Seq Mask-RCNN method, respectively), demonstrating the high segmentation quality in our predictions. Also, TubeFormer-DeepLab employs a smaller clip value ($T = 5$), while other state-of-the-art proposal-based approaches use a large value of clip ($T = 13$ or 36).

*Work done during an internship at Google.

method	backbone	decoder	STQ	SQ	AQ	VPQ
Motion-DeepLab [24]	ResNet-50 + dual ASPP [6]	dual DeepLabv3+ decoder [7] w/ multi-heads	58.0	67.0	51.0	40.0
VPSNet [14]	ResNet-50 + FPN [16]	Mask R-CNN [11] style multi-heads	56.0	61.0	52.0	43.0
TF-DL-B1	ResNet-50 + axial-attention [22]†	tube-transformer	**70.0**	**76.8**	**63.8**	**51.1**

Table 1. **[VPS]** KITTI-STEP *val* set results. †: Axial attention blocks [22] are used in the last two stages.

method	backbone	decoder	mIoU	VC8	VC16
TCB [18]	ResNet-101	spatial-temporal OCRNet [28] + memory aggregation	37.8	87.9	84.0
TF-DL-B1	ResNet-50 + axial-attention [22]†	tube-transformer	**58.0**	**90.1**	**86.8**

Table 2. **[VSS]** VSPW *val* set results. †: Axial attention blocks [22] are used in the last two stages.

method	backbone	decoder	T	AP	AR_1	AR_{10}
MaskProp [2]	ResNet-50 + FPN [16] + HTC [5]	Mask R-CNN [11] style multi-heads w/ mask refinement postprocessing	13	40.0	-	-
	ResNet-101 + FPN [16] + HTC [5]		13	42.5	-	-
	ResNeXt-101 [25] + FPN [16] + HTC [5]		13	44.3	-	-
	ResNeXt-101 [25] + FPN [16] + HTC [5] + deform.STSN [3, 10]		13	46.6	-	-
Seq Mask R-CNN [15]	ResNet-50 + FPN [16]	Mask R-CNN [11] style multi-heads w/ many proposals	36	40.4	41.1	49.7
	ResNet-101 + FPN [16]		36	43.8	46.3	52.6
	ResNeXt-101 [25] + FPN [16]		36	**47.6**	46.3	56.0
VisTR [23]	ResNet-50	DETR [4] style transformer	36	36.2	37.2	42.4
	ResNet-101		36	40.1	38.3	44.9
IFC [13]	ResNet-50 + FPN [16]	DETR [4] style transformer	5	41.0	43.5	52.7
	ResNet-50 + FPN [16]		36	42.8	43.8	51.2
	ResNet-101 + FPN [16]		36	44.6	44.0	52.1
STEm-Seg [1]	ResNet-50 + FPN [16]	3D Conv-based TSE [1] w/ multi-heads	8	30.6	31.6	37.1
	ResNet-101 + FPN [16]		8	34.6	34.4	41.6
TF-DL-B1 (per-pixel)	ResNet-50 + axial-attention [22]†	tube-transformer	5	36.4	40.8	49.5
(per-mask)	ResNet-50 + axial-attention [22]†		5	38.8	44.0	51.4
TF-DL-B4 (per-pixel)	ResNet-50-n4 + axial-attention [22]†		5	45.4	48.3	56.9
(per-mask)	ResNet-50-n4 + axial-attention [22]†		5	47.5	**50.2**	**59.0**

Table 3. **[VIS]** YouTube-VIS-2019 *val* set results. †: Axial attention blocks [22] are used in the last two stages. ResNet-50-n4 scales the number of layers in stage-4 by 4 times (*i.e.*, 24 blocks in total), resulting in a backbone with 104 layers.

2. Visualization

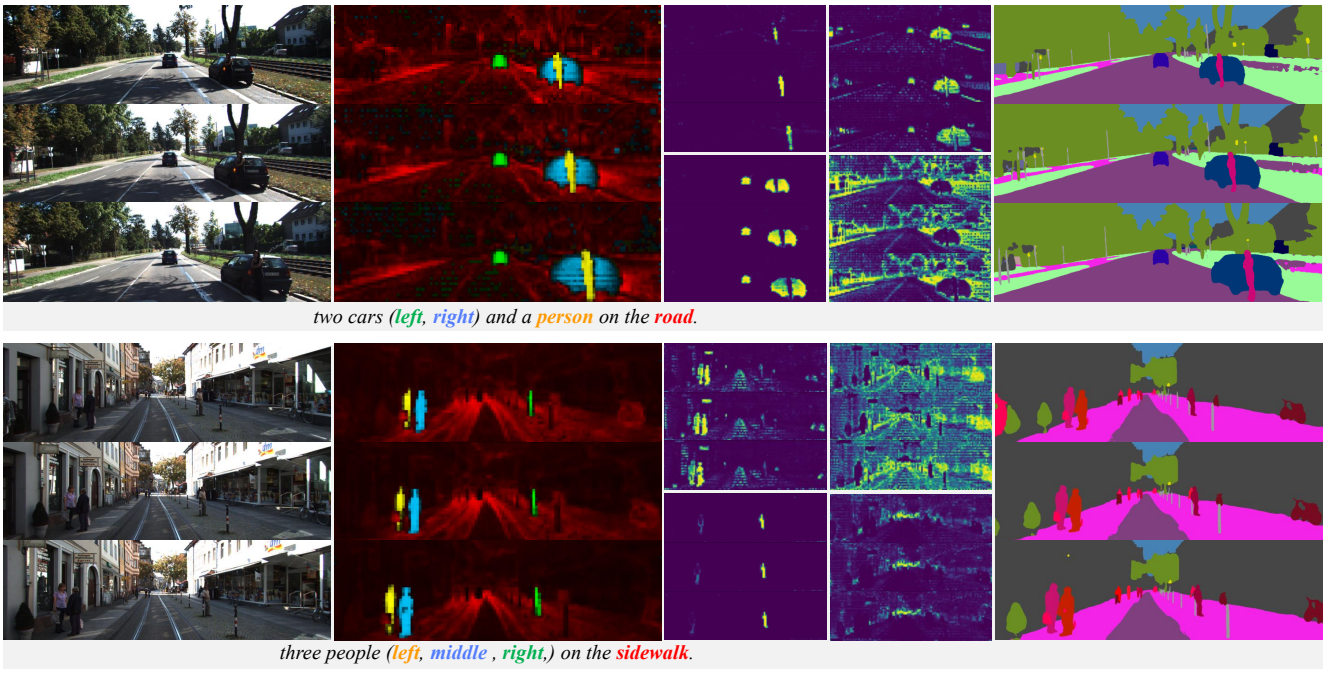
In Fig. 1, 2, and 3, we visualize how the proposed hierarchical dual-path transformer performs attention for video panoptic/semantic/instance segmentation tasks (VPS, VSS, and VIS, respectively). We use input clips of three consecutive frames for visualization. For each sample, we select several output tubes of interest from the TubeFormer-DeepLab prediction. In column-b, we probe the attention weights between the selected tube-specific global memory embeddings and all the pixels. Across all three tasks, we observe the global memory attention is spatio-temporally clustered for individual tube regions, while respecting different requirements among the tasks. That is, one global memory answers for each semantic category in VSS, but for each instance identity in VIS, while both cases appear in VPS task.

In column-c, we select four latent memory indices and visualize their attention maps. Commonly for all tasks, some latent memory learns to spatially specialize on certain areas (left *vs* right side of the scene) or attends to the tube boundaries. Interestingly, we find that some latent memory focuses on relatively far-away region (Fig. 1c-bottom right), which often requires more attention. Sometimes, it has more interests to the moving object parts or small objects (*e.g.*, *moving arms* and *a road-block cone* in Fig. 2c-bottom left and bottom right, respectively).

The task-specific behavior of the latent memory can be also compared between Fig. 2c and Fig. 3c. The latent memory in VSS does not distinguish instances of a same semantic class. In contrast, the attention is instance-specific in VIS. As shown in Fig. 3c-top left, the *occluded noses of two elephants* are highlighted, which is expected to help the instance discrimination. Also, different latent memory attends to a single, or different multiples of the instances.

Additionally, Fig. 4 visualizes our depth-aware video panoptic segmentation results on SemKITTI-DVPS dataset, where TubeFormer-DeepLab is able to generate temporally consistent panoptic segmentation and monocular depth estimation results.

Finally, we attach our *video* prediction results for each dataset in the supplementary materials (see other provided



(a) input frames (b) global memory attention (c) latent memory attention (d) video panoptic segmentation

Figure 1. **[VPS] Visualization on KITTI-STEP sequence**. From left to right: input frames (T=3), global memory attention, latent memory attention, and video *panoptic* segmentation results. The global memory attention is selected for predicted tube regions of interest, and the latent memory attention is selected for 4 (out of L=16) latent memory.

video files).

3. Limitations

Currently, the proposed TubeFormer-DeepLab performs clip-level video segmentation with the clip value $T = 2$ (for VPS and VSS) or $T = 5$ (for VIS). Our model thus performs short-term tracking and may miss objects that have track lengths larger than the used clip value. This limitation is also reflected in the AQ (association quality) reported in Tab. 1 of the main paper (*i.e.*, KITTI-STEP *test* set results). We leave the question about how to efficiently incorporate long-term tracking to TubeFormer-DeepLab for feature work.

In any case, our proposed TubeFormer-DeepLab presents the first attempt to tackle multiple video segmentation tasks from a unified approach. We hope our simple and effective model could serve as a solid baseline for future research.

4. Dataset License

- ImageNet [21]: https://image-net.org/download.php
- Cityscapes [9]: https://www.cityscapes-dataset.com/license/
- COCO [17]: CC BY 4.0
- KITTI-STEP [24]: CC BY-NC-SA 3.0
- VSPW [18]: CC BY 4.0
- Youtube-VIS [26]: CC BY 4.0
- SemKITTI-DVPS [20]: CC BY-NC-SA 4.0

References

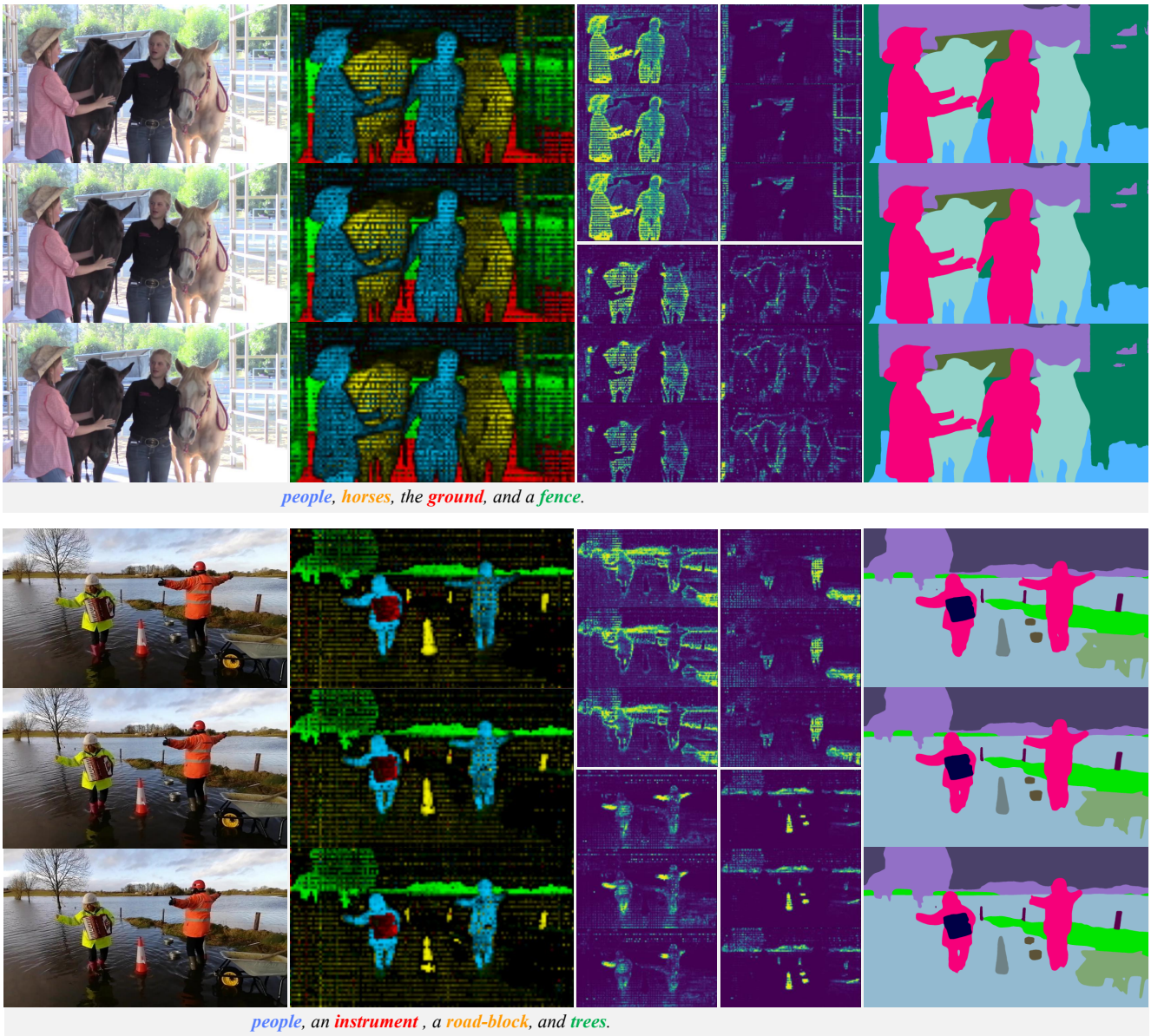
[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 1, 2

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 1, 2

[3] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi,



(a) input frames (b) global memory attention (c) latent memory attention (d) video semantic segmentation

Figure 2. **[VSS] Visualization on VSPW sequence**. From left to right: input frames (T=3), global memory attention, latent memory attention, and video *semantic* segmentation results. The global memory attention is selected for predicted tube regions of interest, and the latent memory attention is selected for 4 (out of L=16) latent memory.

Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2

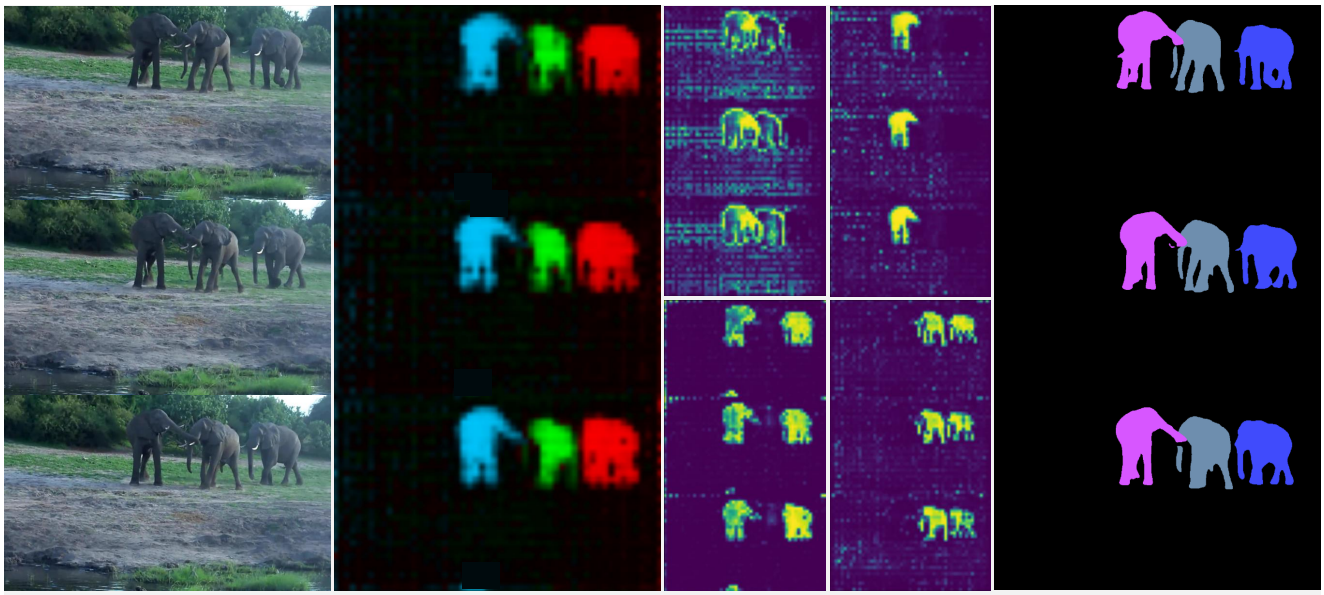
[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 1, 2

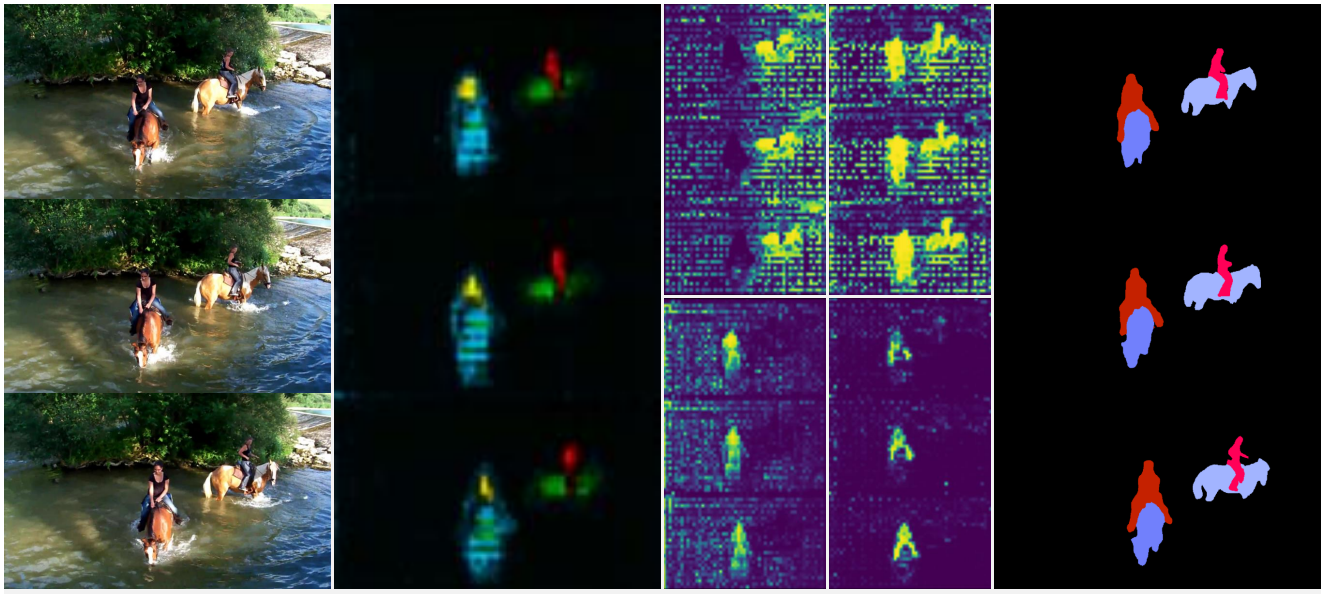
[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 2

[8] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 1

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2





(a) input frames (b) global memory attention (c) latent memory attention (d) video instance segmentation

Figure 3. **[VIS] Visualization on Youtube-VIS 2019 sequence**. From left to right: input frames (T=3), global memory attention, latent memory attention, and video *instance* segmentation results. The global memory attention is selected for predicted tube regions of interest, and the latent memory attention is selected for 4 (out of L=16) latent memory.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2

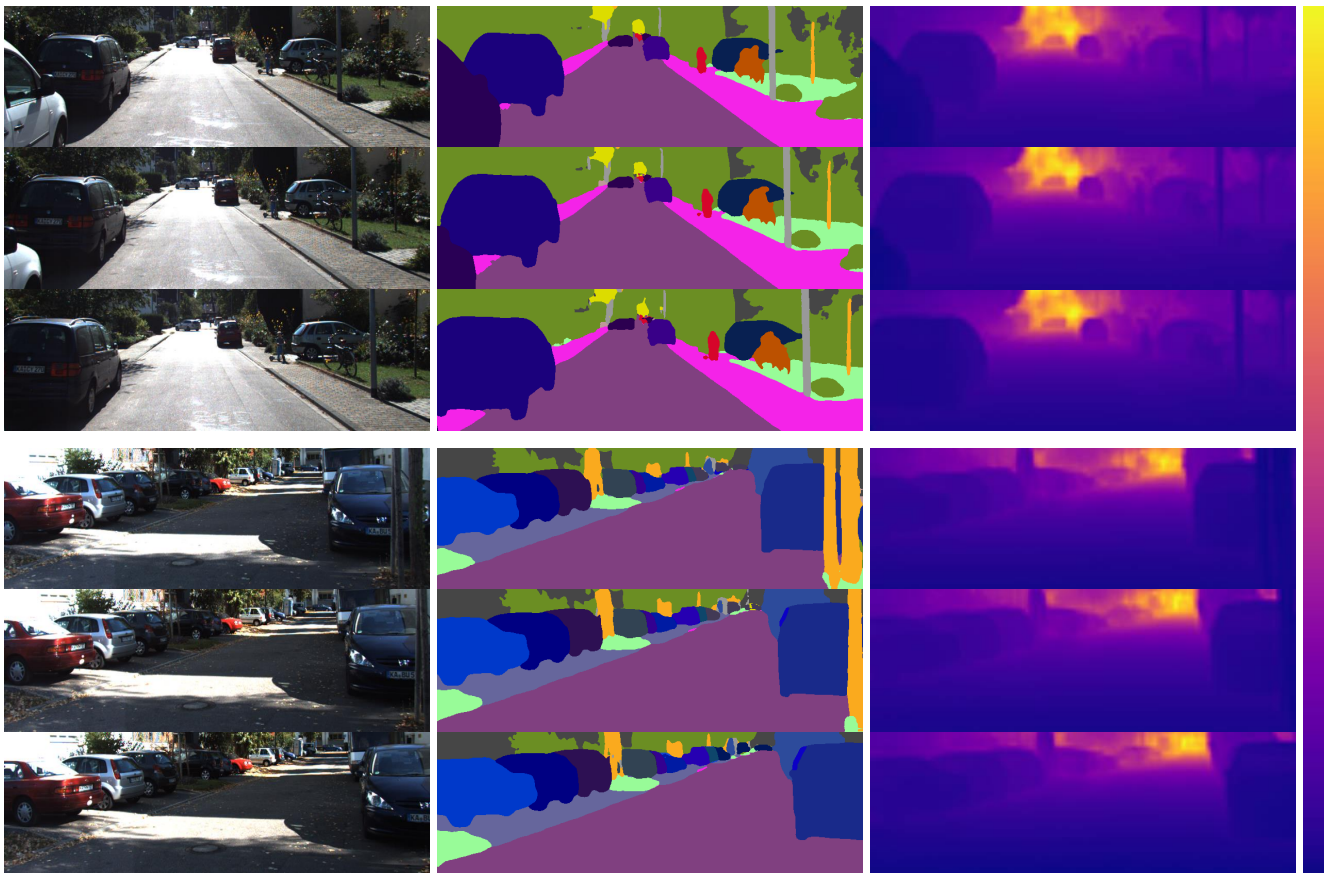
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[13] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In *NeurIPS*, 2021. 1, 2

[14] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 1, 2

[15] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *ICCV*, 2021. 1, 2

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He,



(a) input frames (b) video panoptic segmentation (c) monocular depth prediction

Figure 4. **[DVPS] Visualization on SemKITTI-DVPS sequence**. From left to right: input frames (T=3), video *panoptic* segmentation, and monocular depth prediction results. As the attentions are very similar to those in KITTI-STEP (Fig. 1), here we focus on the depth visualization.

Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 2

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[18] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, 2021. 1, 2, 3

[19] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1

[20] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. ViP-DeepLab: Learning Visual Perception with Depth-aware Video Panoptic Segmentation. In *CVPR*, 2021. 3

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 3

[22] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *ECCV*, 2020. 1, 2

[23] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[24] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixe, and Liang-Chieh Chen. Step: Segmenting and tracking every pixel. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 1, 2, 3

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 2

[26] Linjie Yang, Yuchen Fan, and Ning Xu. Video Instance Segmentation. In *ICCV*, 2019. 3

[27] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille,

and Liang-Chieh Chen. CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation. In *CVPR*, 2022. 1

[28] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 1, 2