PVO: Panoptic Visual Odometry

[Weicai Ye, Xinyue Lan]Co-Authors, Shuo Chen, Yuhang Ming,
Xingyuan Yu, Hujun Bao, Zhaopeng Cui, Guofeng Zhang*

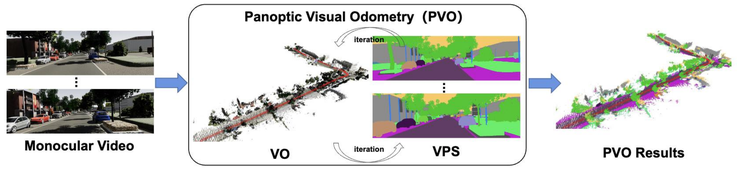
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

zju3dv.github.io/pvo

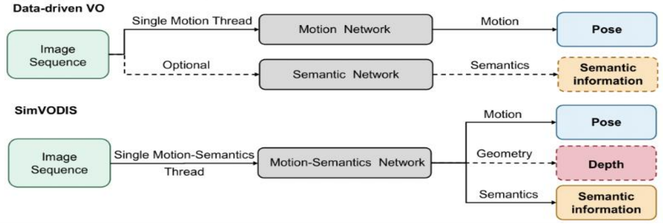
1. Motivation

Our Problem:

- Input: monocular videos
- Output: panoptic 3D map with camera poses

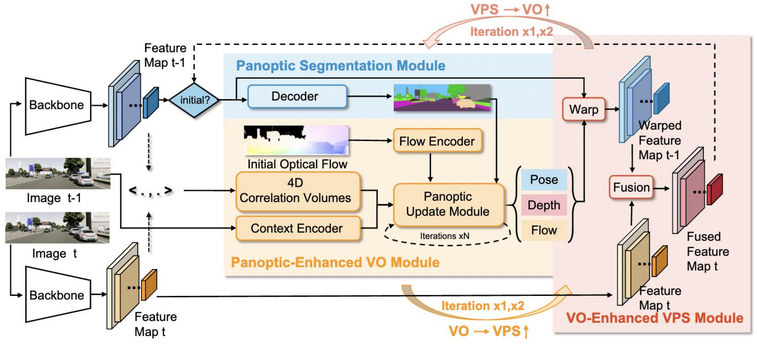


- Existing method: neglect the relevance or contradicting loss



2. Our Solution

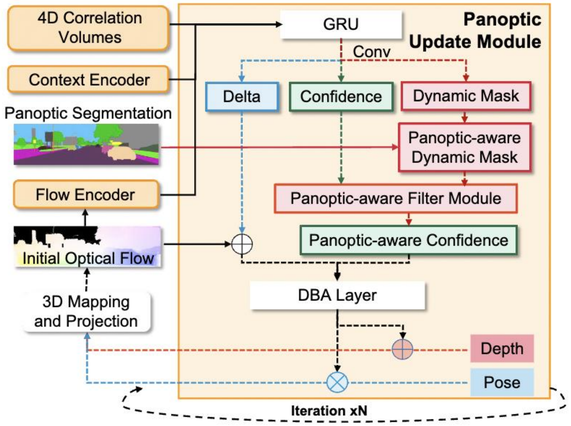
- Unify video panoptic segmentation (VPS) and visual odometry (VO) to model the scene comprehensively
- Mutually beneficial through recurrent iterative optimization



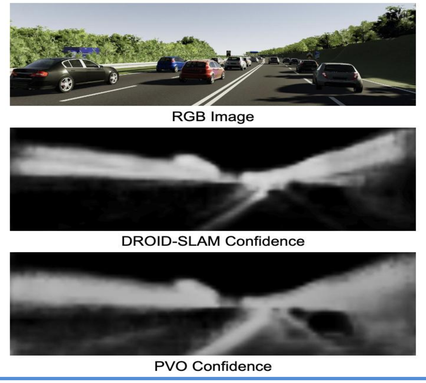
3. Enhanced Module

Panoptic-Enhanced VO Module

- VPS helps VO: adjust weight for stuff and thing

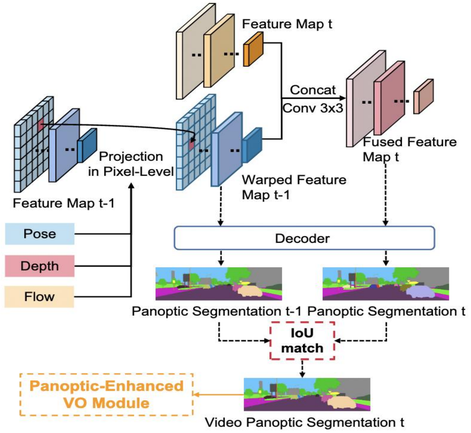


- Panoptic-Aware Confidence: remove the dynamic interference, and keep the static feature



VO-Enhanced VPS Module

- VO helps VPS: tracking from 2D to 3D



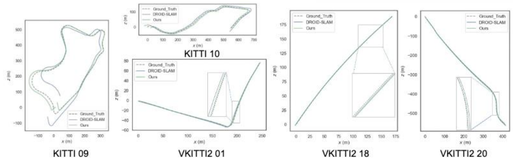
4. Experiments

SOTA VO Results

- Robust pose on dynamic scene VKITTI2

Monocular	01	02	06	18	20	Avg
DROID-SLAM [34]	1.091	**0.025**	0.113	1.156	8.285	2.134
Ours (VPS->VO w/o filter)	0.384	0.061	0.116	0.936	5.375	1.374
Ours (VPS->VO)	0.374	0.057	0.113	0.960	3.487	0.998
Ours (VPS->VO x2)	0.371	0.057	0.113	0.954	3.135	0.926
Ours (VPS->VO x3)	**0.369**	0.055	**0.113**	**0.822**	**3.079**	**0.888**
DROID-SLAM's runtime (FPS)	5.73	12.67	19.96	7.08	10.20	11.13
Ours' runtime (FPS)	4.45	9.69	14.52	6.22	8.10	8.60

- Trajectory Comparison



SOTA VPS Results

- VPS on VKITTI2

Methods on **VKITTI2**	Temporal window size k = 0	k = 5	k = 10	k = 15	VPQ
VPS baseline	58.24 / 60.11 / 57.93	55.50 / 53.78 / 56.28	54.13 / 50.29 / 55.53	53.65 / 48.53 / 55.46	54.90 / 51.95 / 56.05
VPS baseline + w/fusion	59.16 / 67.00 / 56.91	56.27 / 60.98 / 54.96	54.96 / 57.74 / 54.18	54.58 / 55.97 / 54.19	55.81 / 59.23 / 54.85
Ours (VO->VPS + w/o fusion)	58.24 / 60.11 / 57.93	55.67 / 54.44 / 56.28	54.29 / 50.91 / 55.53	53.83 / 49.22 / 55.46	55.04 / 52.48 / 56.05
Ours (VO->VPS + w/fusion + w/o fea loss)	58.51 / 64.07 / 56.97	55.62 / 58.53 / 54.86	54.29 / 55.15 / 54.13	53.94 / 53.40 / 54.19	55.14 / 56.62 / 54.81
Ours (VO->VPS + w/fusion + w/o seg loss)	58.73 / 65.05 / 56.95	55.83 / 59.34 / 54.89	54.51 / 56.01 / 54.15	54.15 / 54.26 / 54.19	55.37 / 57.49 / 54.82
Ours (VO->VPS)	59.18 / 67.00 / 56.94	56.25 / 61.00 / 54.93	54.94 / 57.77 / 54.15	54.57 / 56.01 / 54.17	55.80 / 59.25 / 54.83
Ours (VO->VPS + w/o depth) x2	59.17 / 66.87 / 56.95	56.39 / 61.45 / 56.25	55.04 / 58.15 / 54.15	54.72 / 56.46 / 54.22	55.89 / 59.57 / 54.83
Ours (VO->VPS) x2	**59.18** / 67.00 / 56.94	**56.42** / 61.67 / 54.93	**55.10** / 58.40 / 54.15	**54.84** / 56.67 / 54.17	**55.94** / 59.77 / 54.83

- VPS on Cityscapes

Methods on **Cityscapes-VPS *val***	Temporal window size k = 0	k = 5	k = 10	k = 15	VPQ	FPS
VPSNet-Track	63.1 / 56.4 / 68.0	56.1 / 44.1 / 64.9	53.1 / 39.0 / 63.4	51.3 / 35.4 / 62.9	55.9 / 43.7 / 64.8	4.5
VPSNet-FuseTrack	64.5 / 58.1 / 69.1	57.4 / 45.2 / 66.4	54.1 / 39.5 / 64.7	52.2 / 36.0 / 64.0	57.2 / 44.7 / 66.6	1.3
SiamTrack	64.6 / 58.3 / 69.1	57.6 / 45.6 / 66.6	54.2 / 39.2 / 65.2	52.7 / 36.7 / 64.6	57.3 / 44.7 / 66.4	4.5
PanopticFCN [22] + Ours	**65.6** / 60.0 / 69.7	**57.8** / 45.7 / 66.6	54.3 / 39.5 / 65.1	52.1 / 35.4 / 64.3	**57.5** / 45.1 / 66.4	5.1
VPSNet-FuseTrack + Ours	65.0 / 59.0 / 69.4	57.6 / 45.0 / 66.7	**54.4** / 39.1 / 65.6	**52.8** / 35.8 / 65.2	**57.5** / 44.7 / 66.7	1.1

- VPS on VIPER

Methods on **VIPER**	Temporal window size k = 0	k = 5	k = 10	k = 15	VPQ	FPS
VPSNet-Track	48.1 / 38.0 / 57.1	49.3 / 45.6 / 53.7	45.9 / 37.9 / 52.7	43.2 / 33.6 / 51.6	46.6 / 38.8 / 53.8	5.1
VPSNet-FuseTrack	49.8 / 40.3 / 57.7	51.6 / 49.0 / 53.8	47.2 / 40.4 / 52.8	45.1 / 36.5 / 52.3	48.4 / 41.6 / 53.2	1.6
SiamTrack	51.1 / 42.3 / 58.5	**53.4** / 51.9 / 54.6	49.2 / 44.1 / 53.5	47.2 / 40.3 / 52.9	50.2 / 44.7 / 55.0	5.1
PanopticFCN [22] + Ours	**54.6** / 50.3 / 57.9	51.7 / 44.5 / 57.3	**50.5** / 41.8 / 57.2	**49.1** / 38.9 / 56.9	**51.5** / 43.9 / 57.3	3.6

Video Editing

- Apply PVO to perform motion control, replication, deletion, etc.