PVO: Panoptic Visual Odometry

Supplementary Material

In this supplementary document, we provide more experiment results (Sec. A), such as the ablation study of our method. We further demonstrate the applicability of our method in video editing (Sec. B) and discuss the limitation (Sec. C) of PVO in video editing. We also provide the supplementary video which demonstrates the qualitative results of our method and the video editing effects.

A. Experiments Results

A.1. Ablation Study of Panoptic-Enhanced VO Module

In our Panoptic-Enhanced VO Module, unlike DROID-SLAM [4], we adjust the confidence by incorporating information from the panoptic segmentation. The dynamic mask is adjusted to the panoptic-aware dynamic mask given the initialized panoptic segmentation. Panoptic segmentation treats trees and buildings as stuff (i.e., the background is static), people and cars, etc. as things (i.e., the foreground). So the foreground objects with a high probability of motion are set to dynamic. We show an example of waiting for a traffic light in Fig. A1, where the white color indicates that the parked cars are static. We find the dynamic threshold set as 0.5 may achieve the best results, shown in Tab. A1. The reason is that when the dynamic threshold is small, too many static pixel points may be removed, while the dynamic threshold is too large, and small movements may be ignored. The confidence and panoptic-aware dynamic mask are passed through a panoptic-aware filter module to obtain the panoptic-aware confidence. As shown in Tab. A1, the panoptic-aware filter module can help improve the estimation of camera pose.

We show the qualitative results of the panoptic 3D maps produced by our method, shown in Fig. A2. The supplementary video also shows how our method works.

A.2. Ablation Study of VO-Enhanced VPS Module

Qualitative results in Fig. A3 shows our method can cope with occlusion better on VKITTI2 dataset. Fig. A4 demonstrates our method keeps consistent video panoptic segmentation on Cityscape-VPS dataset, compared with VPSNet-FuseTrack.

B. Video Editing Applications with PVO

In this section, we show the applicability of video editing with PVO, as shown in Fig. B5. We can obtain rich 2D and 3D information from panoptic visual odometry, which can be utilized in video editing.

Fig. B8 illustrates how we can perform consistent video editing using panoptic visual odometry. Firstly, we feed the original video frames from t to t+n into the PVO network. The VO-Enhanced VPS Module and VPS-Enhanced VO Module will get the panoptic segmentation result and optical flow estimation, depth, and pose information for each frame. In addition, the motion of dynamic objects can be decomposed into the dynamic field and static field of the camera. Similarly, the above operation in the new scene can get the whole scene modeling information. We can first select one instance of the original video, then obtain the motion of the target in the new scene by merging the static field of the new scene and the dynamic field of the selected object of the original video, together with additional information such as depth checks and occlusion completion, to complete the video effect of inserting the object into the new scene. In this

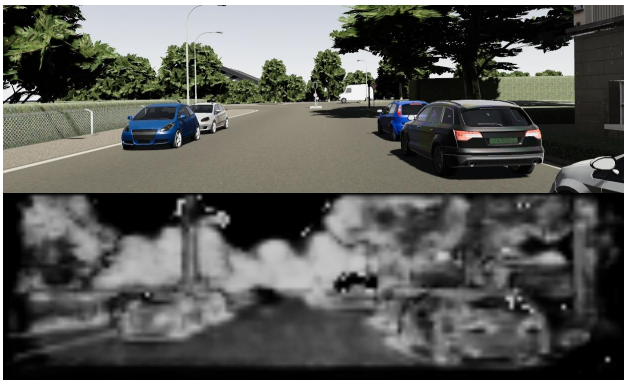


Figure A1. **Dynamic Probability of Parked Cars.** The black color indicates that the confidence tends to be close to 0.

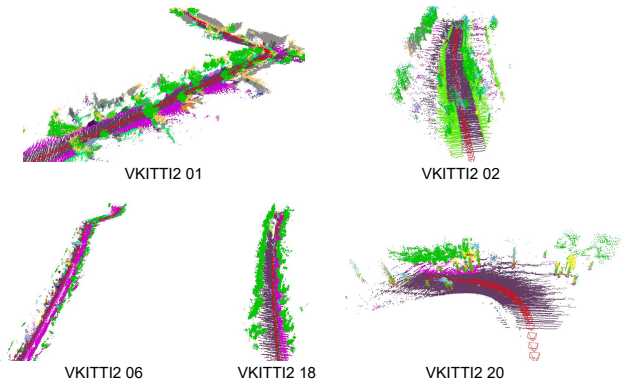


Figure A2. **Qualitative Results of Panoptic 3D Map Produced by PVO on Virtual KITTI Dataset.** We show the panoptic 3D map produced by our method. The red triangles indicate the camera pose, and different colors indicate different instances.

Monocular	vkitti01	vkitti02	vkitti06	vkitti18	vkitti20	Avg
DROID-SLAM	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**
Ours (VPS->VO x3) threshold=0.1	0.377	0.052	**0.112**	0.950	3.240	0.946
Ours (VPS->VO x3) threshold=0.3	0.374	0.054	0.113	0.946	3.107	0.919
Ours (VPS->VO x3) threshold=0.5	0.369	0.055	0.113	0.822	3.079	0.888
Ours (VPS->VO x3) threshold=0.7	0.384	0.059	0.114	0.863	22.993	4.883
Ours (VPS->VO x3) threshold=0.9	1.348	0.065	0.119	0.885	17.337	3.951

Table A1. **Ablation Study of Panoptic-Enhanced VO Module on Virtual KITTI2 Dataset.** Panoptic-Enhanced VO Module outperforms DROID-SLAM on most of the highly dynamic VKITTI2 datasets, and the accuracy of the pose estimation is significantly improved after recurrent iterative optimization. The dynamic threshold set as 0.5 can achieve the best performance.

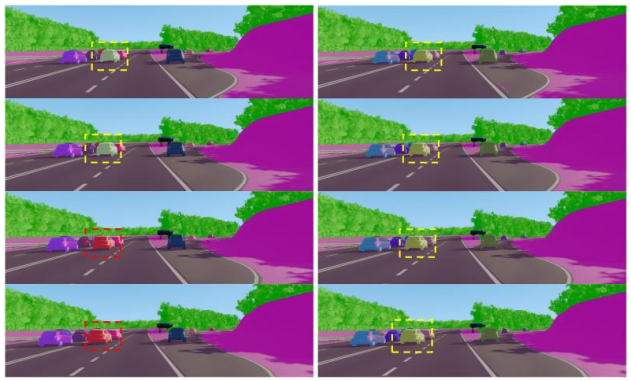


Figure A3. **Comparison Results of Our VO-Enhanced VPS Module with VPS Baseline on VKITTI2 Dataset**. Our method keeps the consistent video segmentation for it is better to cope with occlusion. Different colors indicate tracking failure.

Figure A4. **Comparison Results of Our method with VPSNet-FuseTrack [2] on Cityscape-VPS Val Dataset.** Compared with VPSNet-FuseTrack, our method can keep consistent video segmentation. Different colors indicate tracking failure.

way, we can perform several vivid video effects, including motion control, replication, deletion, and instance interaction. Note that when the initial segmentation is incomplete, with PVO, we can first fill in the occluded parts from multiple views, thus ensuring the integrity of the object, as shown in Fig. B9.

B.1. Ablation Study of Video Editing

We perform an ablation study of PVO in video editing compared with the existing method.

Baseline: We use Video Propagation Network [1] to perform video editing for motion control. The baseline method is generally simple to manipulate objects without considering occlusion, but it doesn't look realistic.

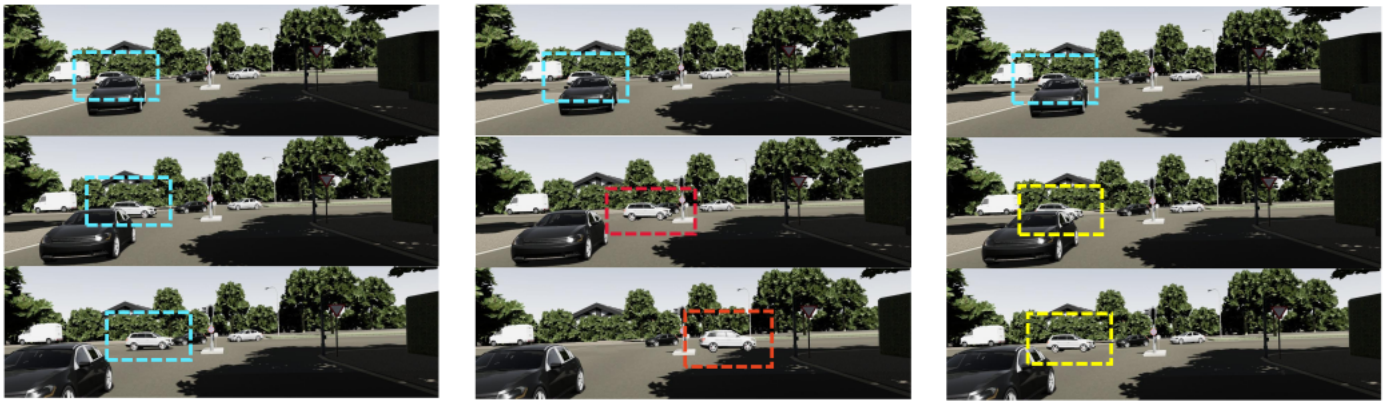
Ours (PVO): The PVO method can better model scene segmentation and motion geometry information and achieve better object manipulation results in occluded scenes, shown in Fig. B10.

B.2. Motion Control

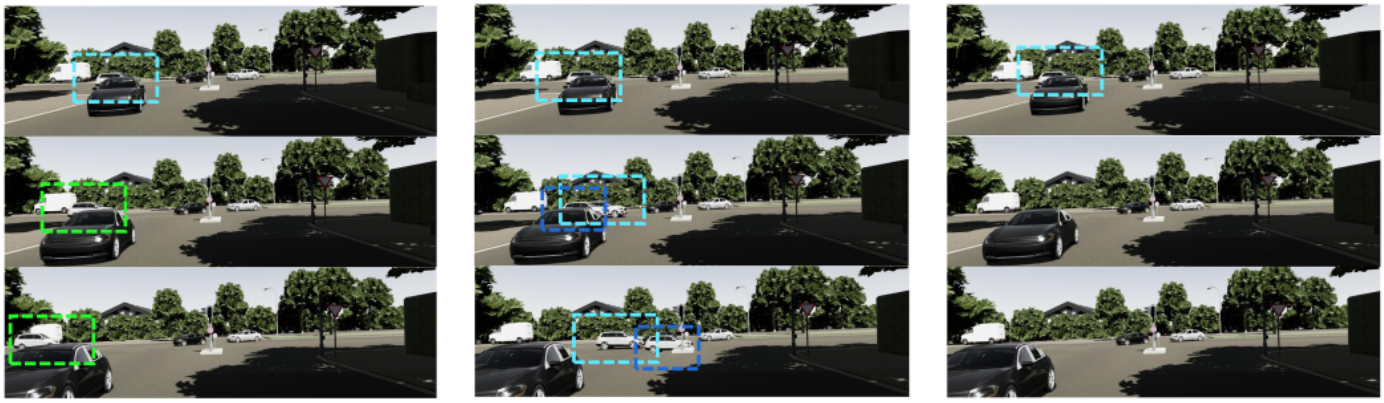
As shown in Fig. B6, we can insert moving objects into the new scene and also directly manipulate the motion patterns of the moving objects of the original video, such as acceleration, deceleration, pause, and rewind. We can also apply PVO to natural scenes such as Cityscapes for motion control, shown in Fig. B7 which shows the generalization of Panoptic Visual Odometry.

B.3. Single vs Multi Instance Interaction

Since PVO provides more comprehensive information about the motion and panoptic segmentation of the scene for the cases such as occlusion, where adjacent frames can



(a) Original video (b) Speed up (c) Slow down



(d) Reverse (e) Copy & paste (f) Delete

Figure B5. **Panoptic Visual Odometry (PVO) can Support Many Video Editing Effects of Motion Control.** With PVO, we can manipulate the white car in the original video with different motions and keep the overall consistency of the video. (a) The car in the original video; (b) Speed the car up; (c) Slow the car down, (d) Put the car in reverse, (e) Copy the new similar car keeping the similar motion, (f) Delete the car. The cyan box indicates the original motion pattern, and the other colors indicate our motion manipulation effect.

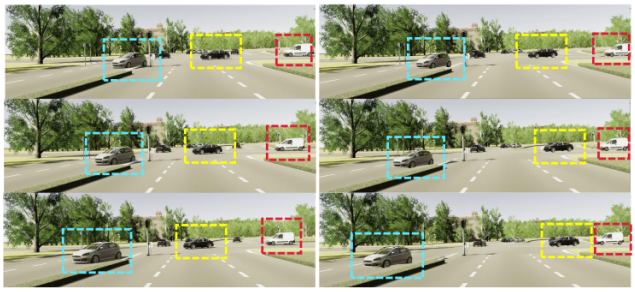


Figure B6. **Multi-Instance Motion Control.** PVO allows a more comprehensive modeling of the scene's motion, panoptic segmentation, and geometric information. PVO can support different motion manipulation of multiple moving objects, even if the camera is also moving. The blue box indicates accelerating the car, the yellow box indicates reversing, and the red box indicates decelerating the car.

be complemented, we can better perform the interaction between different instances, as shown in Fig. B9.

Figure B7. **Generalization Results of Motion Control on Cityscapes Dataset.** PVO demonstrates generalizability in natural scenes.

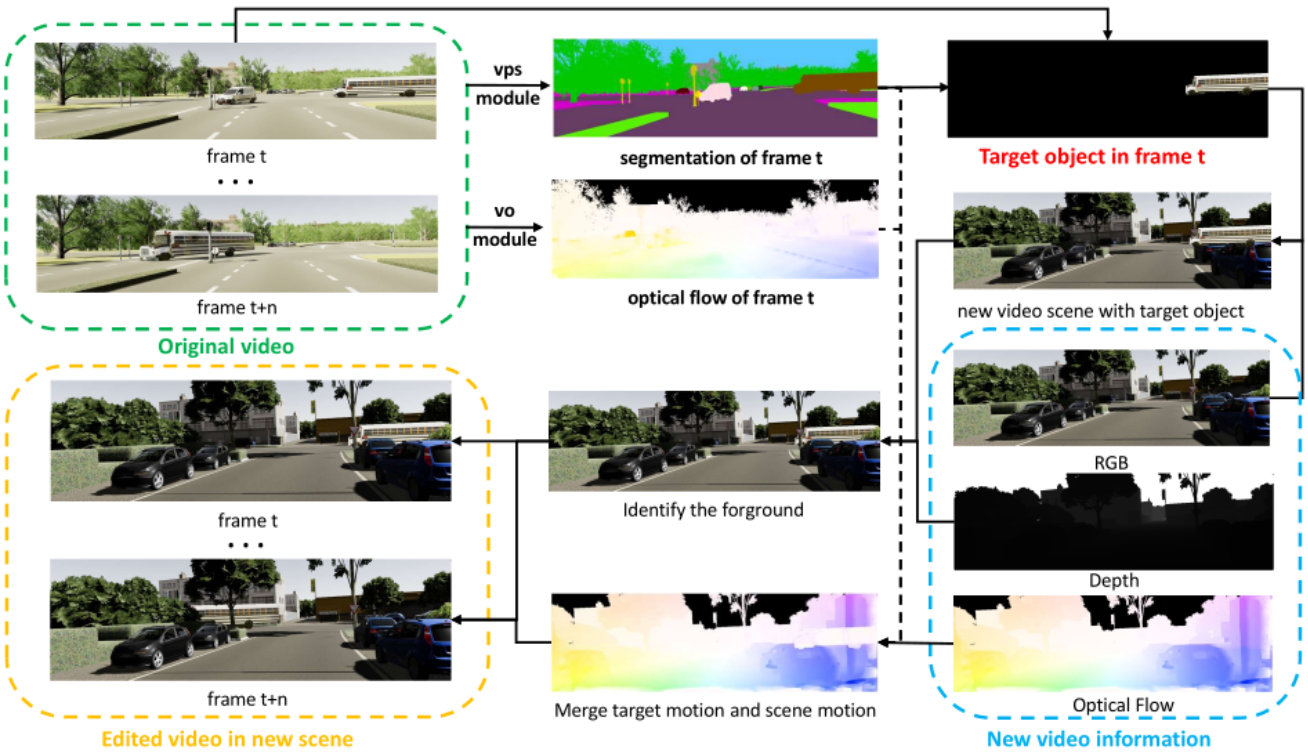
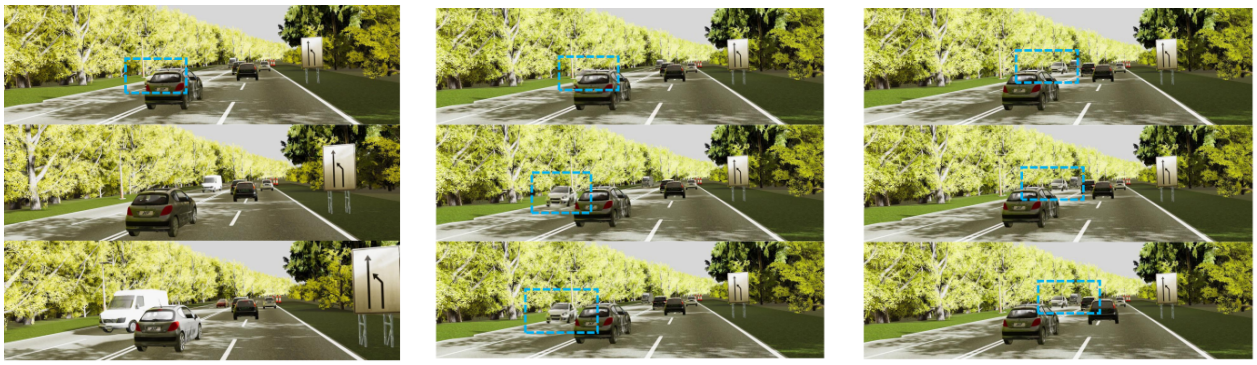


Figure B8. **Video Editing Pipeline with Panoptic Visual Odometry.** PVO comprehensively models the panoptic segmentation and motion information of the entire scene. The motion of dynamic objects can be decomposed into static fields and dynamic fields. We can select an instance and directly manipulate its static fields and dynamic fields to generate a new video. Moreover, the original moving objects can be inserted into the new scene. Some of the occlusion completion and depth checks are taken into account to create a more realistic editing effect.



(a) Original video (b) Pulling cars left and right (c) Pulling cars back and forth

Figure B9. **Multi-Object Occlusion Interaction.** PVO can model the motion, panoptic segmentation, and geometric information of the scene more comprehensively. PVO can support completing multiple mutually occluding moving objects. From left to right, they are a: the original video, b: the occluded part can be restored by pulling the car left and right. c: the occluded part can be restored by pulling the car back and forth.

B.4. Copy and Paste

It's also interesting to copy objects running in other lanes into an empty lane, shown in Fig. B11.

B.5. Delete

If a single lane is overloaded, we can also perform the operation of removing the running vehicle, shown in Fig. B12.

Figure B10. **Ablation study of Video Editing.** Copy the moving objects from the original video to the current moving video. From left to right: the original video, the edited results of the baseline method, and the edited results of the PVO method. PVO can model the complete scene information, such as depth, pose, optical flow, panoptic segmentation, etc. Compared to the baseline method, which edits the scene only by segmentation and optical flow, the results of PVO are more realistic.

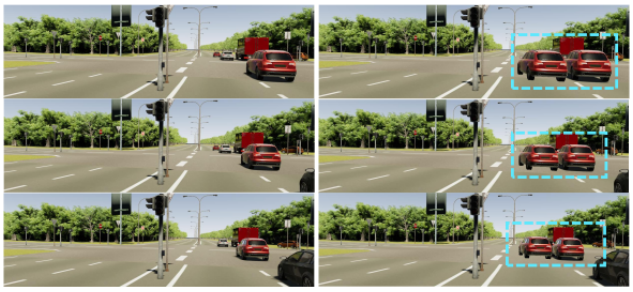


Figure B11. **Copy & Paste.** We can replicate the same moving vehicle in a new lane and keep the video consistent, leveraging the proposed PVO method.

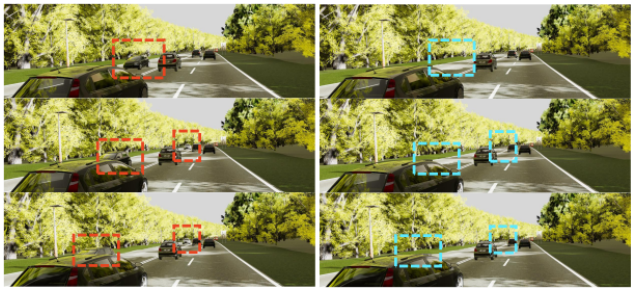


Figure B12. **Delete.** If a single lane is overloaded, we can also perform the operation of removing the moving vehicle using the proposed PVO method. From left to right: the original video, the edited result of removing the car.

C. Discussion

Although PVO can model the panoptic segmentation and motion information of the scene well and support video editing effects such as manipulating the motion patterns of objects. However, it does not take into account the intrinsic physical information of the scene [5], such as lighting, materials, shading, etc., so it cannot make a completely realistic video. In addition, effects [3] related to the objects themselves, such as shadows, cannot be modeled. Fully modeling the movement, panoptic segmentation, effects, and physical properties of the scene is an issue worth exploring. Furthermore, we can explore the application of PVO to autonomous driving simulations to test the robustness of autonomous driving systems by manipulating the motion of objects. We leave this as future work.

References

[1] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video Propagation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 451–461, 2017. 2

[2] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So

Kweon. Video Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020. 2

[3] Erika Lu, Forrester Cole, Tali Dekel, Andrew Zisserman, William T Freeman, and Michael Rubinstein. Omnimatte: Associating Objects and Their Effects in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4507–4515, 2021. 5

[4] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in Neural Information Processing Systems*, 2021. 1

[5] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis. *arXiv preprint arXiv:2210.00647*, 2022. 5