InstMove: Instance Motion for Object-centric Video Segmentation

Qihao Liu*[1] Junfeng Wu*[2] Yi Jiang[3] Xiang Bai[2] Alan Yuille[1] Song Bai[3]
[1]Johns Hopkins University [2]Huazhong University of Science and Technology [3]ByteDance

Appendices

Here we provide implementation details (Sec. 1) and extended experimental results (Sec. 2) omitted from the main paper for brevity.

1. Implementation Details

Training InstMove. We use Adam optimizer with a learning rate of 5×10^{-5} during training. For all experiments, the model is training for 10K iterations on 8 V100 GPUs of 32G RAM, with a batch size of 32. We re-scale all the input image masks to 384×384 with padding to preserve the aspect ratios. We set memory length $l = 256$ and memory size $c = 100$. During training, we randomly select adjacent 3 to 5 frames (the last frame serves as the target frame) to enable the model to handle different input lengths during inference.

2. Comparison with Optical Flow

Optical flow is used to provide motion information in many previous methods. Since it considers pixel-level motion, it can be used to propagate previous object masks to the current frame through a warp layer. In this section, we use RAFT to propagate the object masks and provide a quantitative comparison with our method on the OVIS-Sparse dataset. Specifically, we use flow between frames t and $t-1$ provided by RAFT to propagate the predicted masks $\mathbf{m}_{t-1}$ in the frame $t-1$ to frame t, and then calculate the mask IoUs between the propagated masks and the predicted masks to get the flow score. As the same with the motion score, the flow score is added to the original matching score of VIS methods.

We compare RAFT and our InstMove on the OVIS-Sparse dataset. Two SOTA VIS methods, *i.e*. MinVIS and IDOL, are used. The frames and annotations are kept every 1, 3, 5, or 7 frames (*i.e*. Sparse-1/3/5/7) to simulate different FPS. Note that RAFT is pretrained on a large number of datasets including FlyingChairs [2], FlyingThings [5], FlyingThings3D, Sintel [1], KITTI-2015 [6], and HD1K [4], while the VIS datasets are relatively small, we train our motion model on the OVIS-Sparse training set that only contains 485 videos. As shown in Table 1, our method outperforms the optical flow-based method in different FPS, which demonstrates the robustness and effectiveness of InstMove.

	Sparse-1	Sparse-3	Sparse-5	Sparse-7
MinVIS [3]	19.2	18.9	15.3	15.1
MinVIS + RAFT	20.4	19.6	18.1	16.3
MinVIS + InstMove	**20.8**	**20.0**	**18.2**	**16.7**
IDOL [7]	24.4	21.3	16.5	14.1
IDOL + RAFT	25.7	**21.5**	17.5	15.2
IDOL + InstMove	**27.0**	**21.5**	**18.8**	**16.2**

Table 1. **Effects of instance-level motion module (InstMove) and pixel-level motion module (RAFT) on VIS task.** We report the mAP on the OVIS-Sparse validation set. InstMove outperforms the optical flow-based method in different FPS, which demonstrates the robustness and effectiveness of InstMove. Note that RAFT is pretrained on a large number of datasets while InstMove is only trained on 485 videos.

References

[1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pages 611–625. Springer, 2012. 1

[2] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 1

[3] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 1

[4] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *CVPR*, pages 19–28, 2016. 1

[5] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 1

[6] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 1

[7] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, pages 588–605. Springer, 2022. 1