OCAI: Improving Optical Flow Estimation by Occlusion and Consistency Aware Interpolation

Supplementary Material

1. Implementation and Training Details

1.1. Video Frame Interpolation

Datasets. We use Sintel and KITTI datasets, which are standard Optical Flow datasets. Sintel (clean) dataset consists of 20 $\sim$ 50 consecutive frames in 23 Videos. In 12 FPS $\rightarrow$ 24 FPS frame interpolation, we load three consecutive frames (I_1, I_2, I_3), and use I_1 and I_3 as an input and generate $\hat{I}_2$ image. Then, we compute the PSNR, SSIM, and LPIPS (using AlexNet and VGG) metrics. And then, we load next frames (I_2, I_3, I_4), and generate $\hat{I}_3$ using I_2 and I_4 frames. We generate all frames from $\hat{I}_2$ to $\hat{I}_{N-1}$ images. Here, N is the number of the frame in each video clip (total 1018 pairs). In 6 FPS $\rightarrow$ 12 FPS frame interpolation, we load (I_1, I_3, I_5), and generate $\hat{I}_3$. Then, we generate frames from $\hat{I}_3$ to $\hat{I}_{N-2}$ (total 972 pairs). KITTI (multiview) train dataset consists of 21 consecutive frames in 200 videos. We generate $\hat{I}_2$ to $\hat{I}_{20}$ frames, and there are 3800 pairs.

Implementation Details. We use IFRNet [5], VFIFormer [8], RIFE [2], EMA-VFI [11], and AMT [6] VFI algorithms as our backward warping baselines. We use their official codes and weights trained on Vimeo90k.[1] We also use Soft-Splatting [9] and RIPR of RealFlow [1] algorithm as our forward warping baselines. We use official codes, Vimeo trained weight for Soft-Splatting, and FlyingChairs+FlyingThings3D trained weight for RIPR.[2] RIPR and our OCAI use RAFT [10] optical flow model, and we also use the same weight with RIPR for fair comparison. We set α in Eq. 10 to 50. Higher α shows good performance as shown in Table 1. However, when it is set too high, e.g., above 100, the result becomes *not a number*.

Table 1. Video Frame Interpolation results on KITTI. We evaluate the VFI with different α weights.

α	PSNR / SSIM ↑	LPIPS (A) / (V) ↓
1	21.98 / 0.756	0.114 / 0.192
10	22.06 / 0.757	0.112 / 0.191
50	**22.08 / 0.758**	**0.112 / 0.190**
100	NA / NA	NA / NA

1.2. Optical Flow

Dataset. We follow semi-supervised optical flow training settings from previous work, e.g., FlowSupervisor [3], RealFlow [1], and DistractFlow [4].In Sintel test evaluation, we follow DistractFlow training pipeline and use Sintel training dataset and Monkaa dataset. In KITTI test evaluation, FlowSupervisor and DistractFlow use additional unlabeled datasets such as Driving and Spring, but RealFlow uses only KITTI multi-view training dataset. In our experiment, we follow RealFlow and use only KITTI multi-view training dataset.

Implementation Details. We follow FlowSupervisor, RealFlow, and DistractFlow settings. We set τ and w as 0.95 and 1 in Eq. 13 and 14, same as in DistractFlow. We use initial decay rate in EMA of 0.99 and gradually increase it to 0.9996. Since our optical flow model already has been trained on C+T in a semi-supervised setting, we use a higher initial decay rate compared to [7] and use the same terminal decay rate as [7].

2. Additional Video Frame Interpolation results

We generate more inter-frame images in Fig. 1, 2 on KITTI and Sintel datasets. In addition, we also generate more inter-frames with different t values (t = 0.2, 0.4, 0.6, 0.8). Since backward warping based VFI algorithms cannot generate continuous I_t images, we compare inter-frames generated by our OCAI and RIPR from RealFlow in Fig. 3.

[1] IFRNet : https://github.com/ltkong218/IFRNet, VFI-Former: https://github.com/dvlab-research/VFIformer, RIFT: https://github.com/megvii-research/ECCV2022-RIFE, EMA-VFI: https://github.com/MCG-NJU/EMA-VFI, AMT: https://github.com/MCG-NKU/AMT

[2] Soft-Splatting: https://github.com/sniklaus/softmax-splatting RealFlow: https://github.com/megvii-research/RealFlow

References

[1] Yunhui Han, Kunming Luo, Ao Luo, Jiangyu Liu, Haoqiang Fan, Guiming Luo, and Shuaicheng Liu. Realflow: Em-based realistic optical flow dataset generation from videos. *arXiv preprint arXiv:2207.11075*, 2022. 1, 3, 4

[2] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 1

[3] Woobin Im, Sebin Lee, and Sung-Eui Yoon. Semi-supervised learning of optical flow by flow supervisor. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[4] Jisoo Jeong, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Distractflow: Improving optical flow estimation via realistic distractions and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13700, 2023. 1

[5] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 1, 3, 4

[6] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 1, 3, 4

[7] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1

[8] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 1, 3, 4

[9] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020. 1

[10] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1

[11] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023. 1, 3, 4

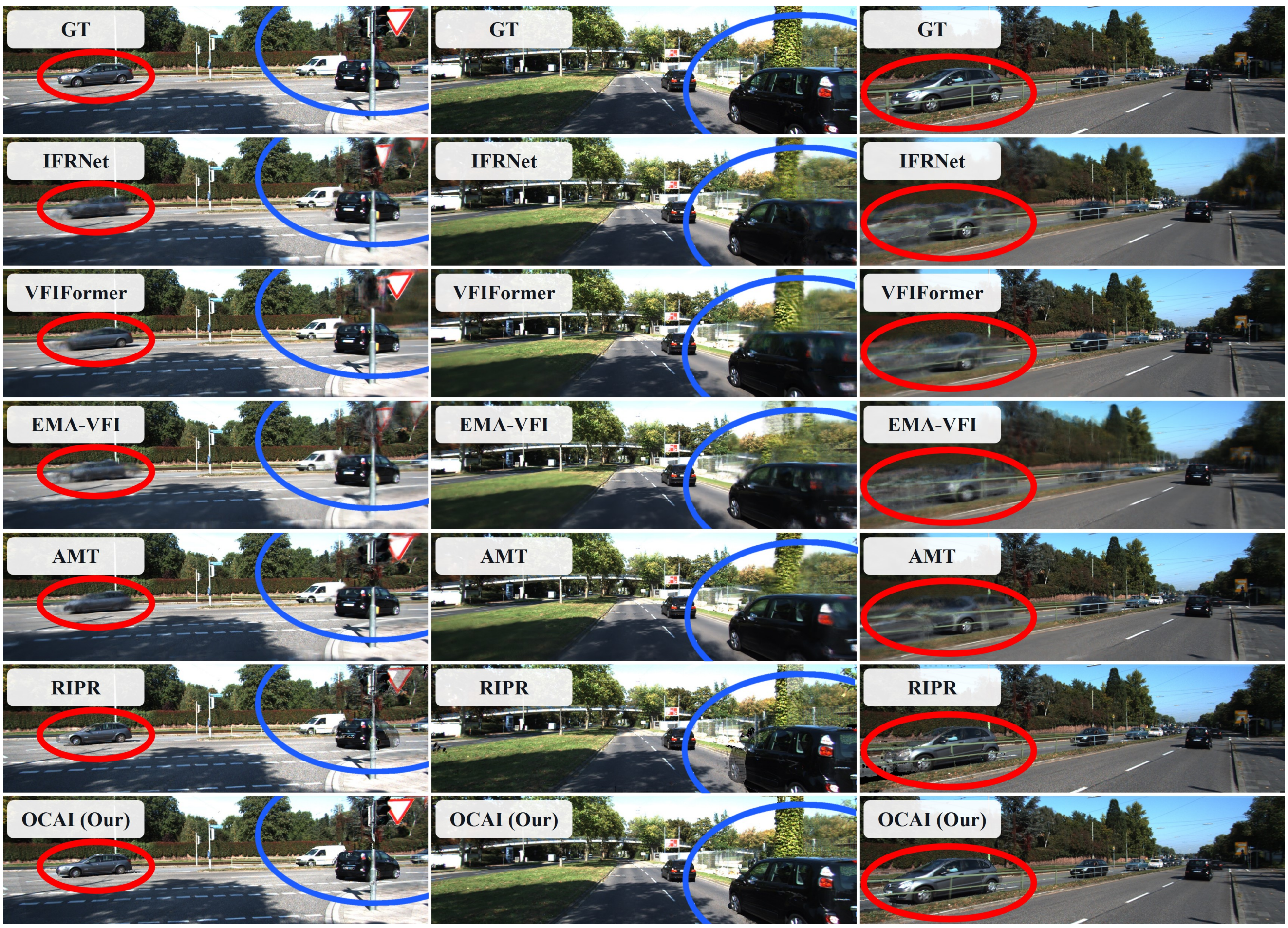

Figure 1. Video Frame Interpolation (VFI) results on KITTI. First row is the ground truth. Second to fifth rows are outputs of SOTA VFI models [5, 6, 8, 11]. Sixth row is the output of using RealFlow [1] for VFI. Bottom row shows our OCAI results.

Figure 2. Video Frame Interpolation (VFI) results on Sintel (clean). First row is the ground truth. Second to fifth rows are outputs of SOTA VFI models [5, 6, 8, 11]. Sixth row is the output of using RealFlow [1] for VFI. Bottom row shows our OCAI results.

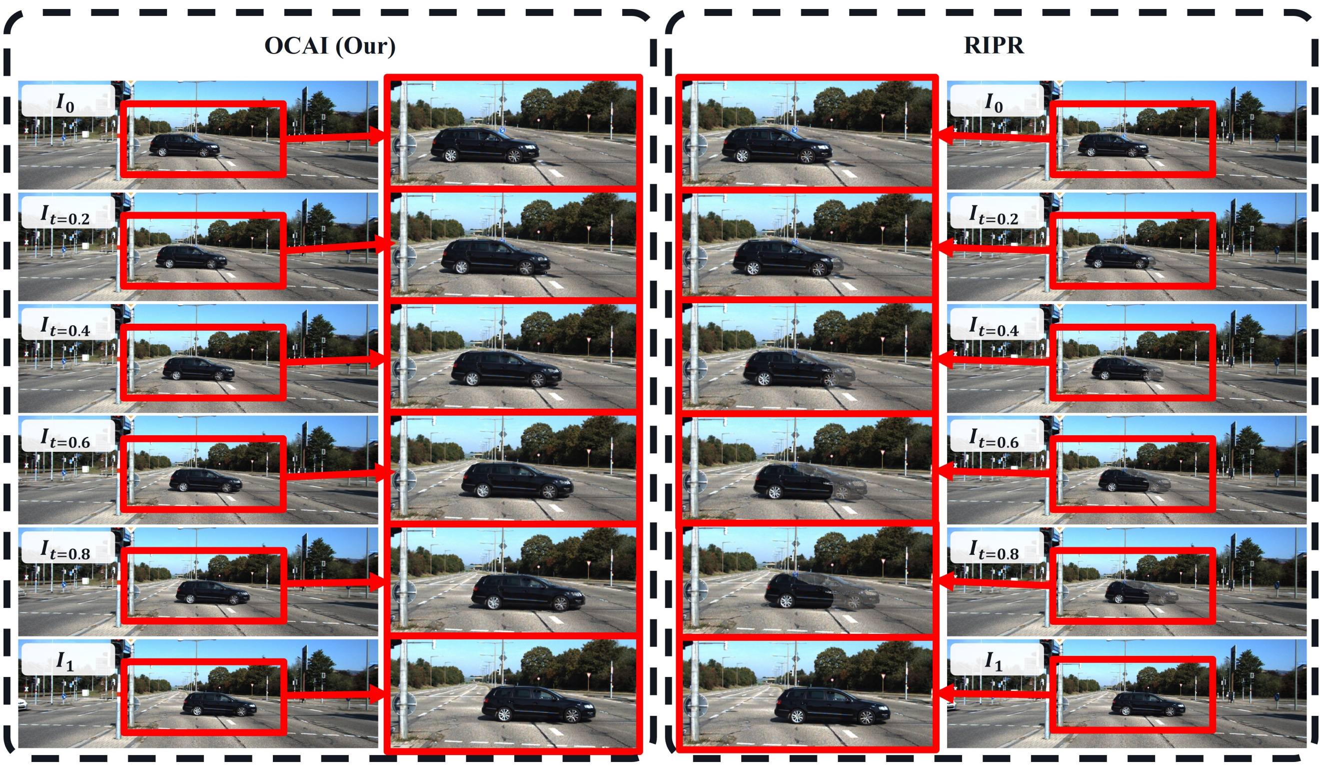

Figure 3. Video Frame Interpolation (VFI) results on KITTI. We generate different I_t images (for $t = 0.2,\ 0.4,\ 0.6,\ 0.8$). Since backward warping cannot generate continuous inter-frames, we generate results using RIPR from RealFlow and our proposed OCAI approach.