RealFlow: EM-based Realistic Optical Flow Dataset Generation from Videos (Supplementary Materials)

Yunhui Han[1], Kunming Luo[2], Ao Luo[2], Jiangyu Liu[2], Haoqiang Fan[2], Guiming Luo[1], and Shuaicheng Liu[3,2,†]

[1] School of Software, Tsinghua University, Beijing 100084, China
{hanyh19@mails.tsinghua.edu.cn, gluo@tsinghua.edu.cn}
[2] Megvii Technology, Beijing, China
{luokunming,luoao02,liujiangyu,fhq}@megvii.com
[3] University of Electronic Science and Technology of China, Chengdu, China
liushuaicheng@uestc.edu.cn
[†]Corresponding Author

To make our RealFlow self-contained, we provide more details in this document including: 1) detailed information of the datasets 2) experiments of the dataset quantity. 3) results of splatting and hole filling. 4) the efficiency of generation. 5) limitation of our method. 6) more qualitative comparison results and visualization of RealFlow samples from various videos.

1 Datasets Detail

In this section, we describe the detailed information of relevant datasets and the generated datasets of our RealFlow. For convenience, we summarize our generated datasets in Table.1.

Flying Chairs [4] and **Flying Things** [17]**:** These two synthetic datasets are generated by randomly moving foreground objects on top of a background image. State-of-the-art supervised networks usually train on Chairs and Things.

Virtual KITTI [5]**:** Virtual KITTI is a synthetic dataset, which contains 50 high-resolution videos generated from 5 different virtual urban environments under different weather conditions. These environments are created by the Unity engine and the flow labels are automatically generated.

DAVIS [3]**:** DVAIS dataset consists of high-quality video sequences under various kinds of scenes, which have been widely used for video object segmentation. It only provides densely annotated segmentation labels but not optical flow labels. We use 10,581 images from DAVIS challenge 2019 to generate **RF-DAVIS**.

ALOV [13] **and BDD100K** [18]**:** ALOV dataset is a diverse real-world video sequence for tracking, ranging from easy to difficult, which covers as diverse circumstances as possible. This dataset consists of 315 video sequences and we capture 75,581 image pairs from them. BDD100K dataset is a large-scale diverse driving video database. It consists of 100,000 videos covering different weather conditions and urban scenes, ranging from daytime to nighttime. We sampled part of the videos and capture 86,128 images pairs. There is no flow label for

Table 1: An overview of our created datasets.

Dataset	Quantity	Source Data
RF-Ktrain	4,000	KITTI-multiview(training)
RF-Ktest	3,989	KITTI-multiview(testing)
RF-Sintel	1,041	Sintel-Clean(training)
RF-DAVIS	10,581	DAVIS-2019 video sequence
RF-AB	161,709	ALOV and BDD100K video sequence

Table 2: Analyze the impact of the dataset quantity. We use the datasets with different amounts of training pairs to train RAFT with the same implementation. The learned models are evaluated on KITTI 2015 train set and Sintel Final train set, where the end-point error (epe) is used as the evaluation metric.

Dataset	Quantity	KITTI15	Sintel F.
dCOCO[1]	100k	3.81	3.90
RF-AB	10k	3.83	3.51
RF-AB	40k	3.65	3.35
RF-AB	80k	3.60	3.32
RF-AB	160k	3.48	3.28

these image pairs, so we use RealFlow to create a large diverse real-world dataset with flow label, named **RF-AB**.

KITTI [6,11]: KITTI2012 and KITTI2015 dataset takes advantage of autonomous driving platform to develop challenging benchmarks for the tasks of optical flow estimation. They provide around 200 training pairs and 200 test pairs. There are multi-view extensions (4,000 training and 3,989 testing) datasets with no ground truth. We use the multi-view extension videos(training and testing) of KITTI 2015 to generate new datasets **RF-Ktrain** and **RF-Ktest**.

Sintel [2]: Sintel is a synthetic flow benchmark derived from the open-source 3D animated short film, which contains 1,041 training pairs and 564 test pairs. 'Clean' and 'Final' rendering sets are provided. We use the images from the training set to generate our **RF-Sintel**.

2 Dataset Quantity

In this section, we analyze the impact of different amounts of datasets for network training. Specifically, we reduce the amount of our RF-AB dataset to half, one quarter, and one-sixteenth, respectively. Then we train RAFT from scratch on these datasets and evaluate the learned models on the train sets of KITTI 2015 and Sintel Final. As shown in Table. 2, a better supervised network can be obtained by training on a larger dataset, which demonstrates that the quantity of labeled pairs is a crucial characteristic of optical flow dataset. Moreover, we also compare our datasets with dCOCO [1]. We can notice that training on our RF-AB dataset with only $10,000$ pairs can achieve a better result than on dCOCO, which contains $100,000$ training pairs. This phenomenon further proves that the realism of motion is a crucial factor for dataset generation.

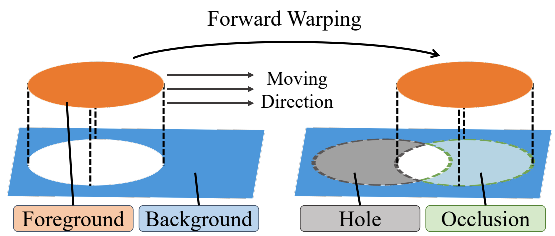


Fig. 1: A toy example of how the hole and occlusion problems occur in the forward warping process. The reference image (left) is divided into two regions: the foreground object (the orange ellipse) and the background (the blue plane). The foreground object is moving to the right. When applying forward warping, the hole problem occurs in the gray region because no pixel is assigned to this area. The occlusion problem occurs in the green region because multiple pixels are assigned to this area.

3 Splatting and Hole Filling

In our Realistic Image Pair Rendering (RIPR) method, we first forward-warp the reference image to the target view and then use splatting and our Bi-directional Hole Filling (BHF) to address the occlusion and hole problems. Here, we first present how the occlusion and hole problems occur in the forward warping process. Then, we show some visual results to analyze the details of the splatting and hole filling methods.

Fig. 1 is a toy example to illustrate how the hole and occlusion problems occur. We divide the reference image into the foreground region (the orange ellipse) and background region (the blue plane), where the foreground region is moving to the right in the next view and the background remains stationary. When the reference image is forward-warped to the next view, the foreground pixels and some background pixels are assigned to the same place, thus the occlusion problem occurs as the green region highlights. Besides, the hole problem occurs when no pixel is assigned to some areas as indicated by the gray region in Fig. 1.

For splatting, as mentioned in our main text, max splatting renders more realistic images than softmax splatting. However, we notice that softmax splatting can produce better dataset than max splatting in our ablation study. Here, we explain why this happens by an example from Sintel dataset that is shown in Fig. 2. We use softmax splatting and max splatting to warp the reference image to the target view based on ground truth optical flow and different depth maps. In Fig. 2, the first line shows the original target image 'Real Image2' and the generated target image 'Image2 Max-GTD' by using max splatting based on the ground truth depth. The second line shows the estimated depth by DPT [12] and the ground truth depth. The third line shows the softmax splatting result and max splatting result based on the estimated depth. As can be seen from Fig. 2, when the depth map is correct (using ground truth depth or correct depth esti-

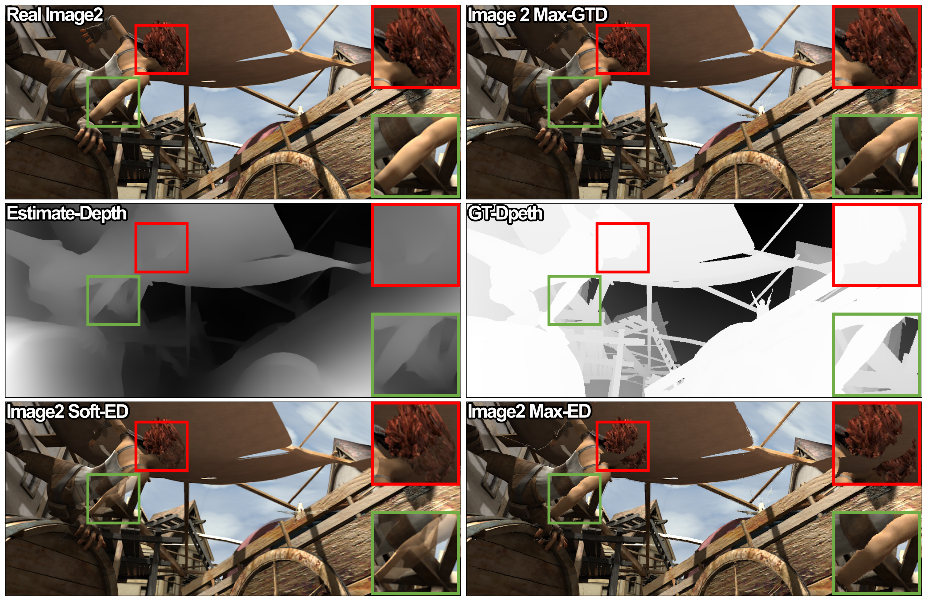


Fig. 2: An example to show softmax splatting and max splatting with ground truth depth and the estimated depth by DPT [12]. Zoom-in patches are shown in the right side. From top to bottom: the original image2 and the generated image2 by max splatting with ground truth depth, the estimated depth and ground truth depth, the softmax splatting result and max splatting result based on the estimated depth. We can notice that when the depth is correct, using max splatting can obtain realistic rendering result, as shown by the green boxes. However, when depth is incorrect, using max splatting may cause harmful tearing of the texture. This problem can be alleviated by using softmax splatting, as shown by the red boxes.

mation), max splatting method can generate a realistic result, which is almost the same as the original target image. However, as highlighted by the red box, the incorrect depth may cause harmful tearing of the texture in the max splatting result, which may reduce the quality of the dataset generated by the max splatting method. This problem can be alleviated by using softmax splatting method, which can obtain a translucent fusion result in these occlusion areas so that the information of the foreground objects can be partially provided even with incorrect depth estimation. This is the reason why sofmax splatting can produce better results than the max version, though max splatting can provide image pairs look more realistic.

For hole filling, in ablation study, we have validated that our BHF outperforms the deep neural inpainting method RFR [7]. Fig. 3 presents some examples for qualitative comparison. For the fair comparison, we use the RFR [7] pre-trained on the large-scale outdoor dataset PARIS, and fine-tuned on KITTI

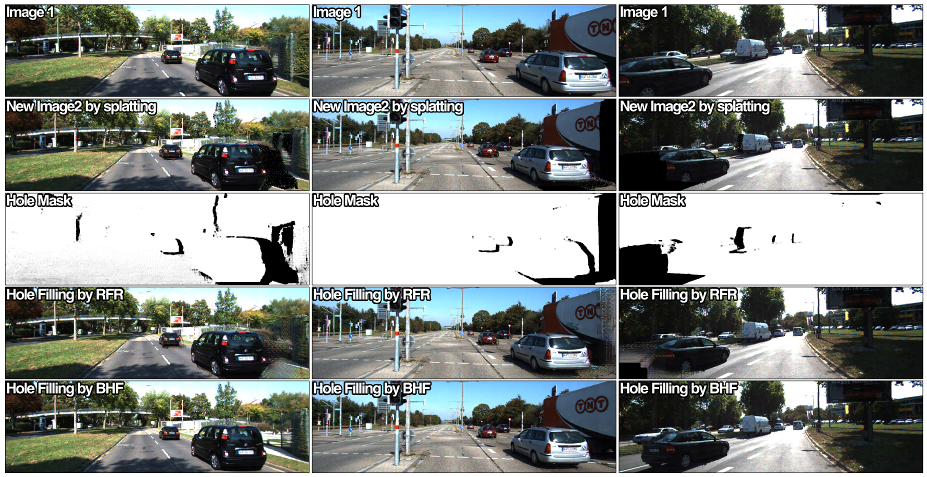


Fig. 3: Qualitative comparison of different methods for hole filling. Top to bottom: the reference images, the 'new image 2' generated by splatting, the hole masks, hole filling results by RFR [7], and hole filling results by our BHF method.

dataset with 100k steps to inpaint these holes. As can be seen, inpainting method RFR may introduce artifacts in these hole regions, while our BHF method can produce realistic image for dataset construction.

4 Efficiency of RealFlow

The efficiency is an essential indicator of a dataset generation method. Traditional manually annotation method [8] may cost 20 $\sim$ 30 minutes per image pair, especially for challenging scenes. AutoFlow[14] takes about 7 days to finish 8 searching iterations using 48 NVIDIA P100 GPUs for the rendering hyperparameters searching and then use the learned hyperparamters to render image pairs. In contrast, our method can automatically generate a large amount of training pairs with high efficiency. During the E-step, the running time of our RealFlow is 0.53s for generating a training pair with a resolution of 512 $\times$ 960 using only 1 NVIDIA 2080Ti GPU. The time consumption of the M-step in our RealFlow depends on the training of the deep neural network. In this work, we train RAFT [15] for 120k iterations with a batch size of 5 on 1 NVIDIA 2080Ti GPU. Using more computational resources may reduce the time consumption of our RealFlow.

5 Limitation

In Fig. 4, as illustrated by the red circle, there is discontinuous illumination artifact in the generated new image 2. This is because that the brightness of the

Fig. 4: Limitation of our RealFlow. As pointed by the red circle, there is discontinuous illumination artifact in the generated new image 2. This is because that the brightness of the car changes greatly in the original image pair.

car changes greatly in the original image pair. During the hole filling process of our RealFlow, object regions with different brightness are used to fill the occlusion regions.

6 Qualitative Results

6.1 Comparison with Dataset Generation Methods

In Fig. 5, we show example training pairs of our RealFlow, Depthstillation [1], AutoFlow [14] and FlyingThings [17]. As can be seen, the training pair of Depthstillation contains a large region of artifacts that reduce the realism of the image. AutoFlow and FlyingThings samples are generated by moving the foreground image patches or objects on the background image, thus the scene objects and their motions are synthesized and can not match the real-world scene. Compared with these methods, the motion and image content of our RealFlow is realistic because the training pair is generated from real-world videos.

The statistics of motion magnitude for different datasets have been shown in Fig. 6. Our RF-AB dataset exhibits an exponential falloff like Sintel and Flyingchairs but is more smooth. The motion of KITTI mainly concentrates in small range which forms a steep polyline. AutoFlow has few small motions and focuses on middle-range motions. The statistics of motion magnitude for RF-AB is similar to the KITTI, when compared with flyingchairs and Autoflow, which maybe one of the factors that lead to the effectiveness when evaluating on KITTI.

We also provide qualitative comparison of our RealFlow with previous dataset generation method Depthstillation [1] in Fig. 7. Specifically, Depthstillation uses DAVIS and KITTI multi-view test videos to generate optical flow training

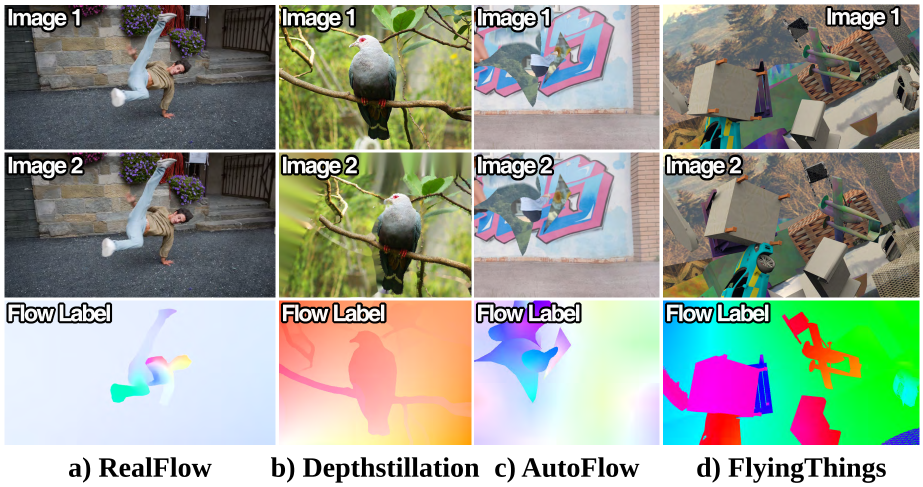


Fig. 5: Example training pairs of our RealFlow and other dataset generation methods, such as Depthstillation [1], AutoFlow [14] and FlyingThings [17].

dataset dDAVIS and dKITTI. For fair comparison, we also use the same source videos to generate RF-DAVIS and RF-Ktest datasets. Then we train RAFT on these datasets with the same hyperparameters and do evaluation on KITTI 2015 train set. The error map of each sample is also visualized with the EPE error depicted in the top-right corner. As can be seen from Fig. 7, models trained on our RF-DAVIS and RF-Ktest datasets can produce better optical flow estimation results than trained on dDAVIS and dKITTI, which demonstrates the effectiveness of our dataset generation method RealFlow.

6.2 Comparison with Unsupervised Methods

There is a set of unsupervised methods that can learn deep neural networks using only video sequences without optical flow labels. We also compare our RealFlow with state-of-the-art unsupervised methods. We first generate an optical flow dataset RF-Ktrain using KITTI multi-view train set, which is also used as the training set of the unsupervised methods. Then we train RAFT on our RF-Ktrain and do evaluation on KITTI 2015 train set. We show qualitative comparison result in Fig. 8, where 'C+T' is the baseline model trained on synthetic datasets FlyingChairs and FlyingThings, UPFlow [9] is the state-of-the-art unsupervised method that trained on the same source videos mentioned above. The End-Point-Error (EPE) is used as the evaluation metric and the error maps are also visualized, where correct predictions are depicted in blue and wrong ones in red. As can be seen, optical estimation network trained on our generated dataset can produce better results than unsupervised methods.

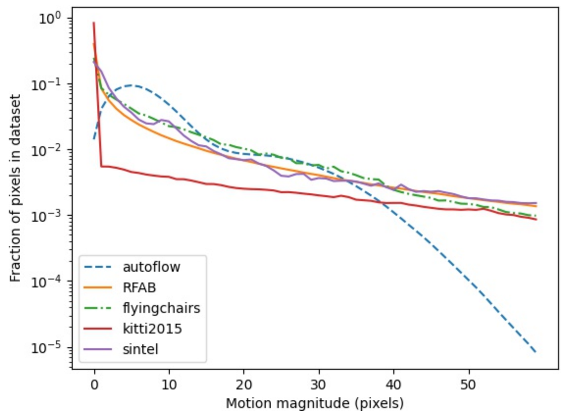


Fig. 6: Histogram of motion magnitude for the generated dataset RFAB and existing datasets.

6.3 Comparison with Supervised Methods

The original RAFT [15] is first pre-trained on synthetic datasets FlyingChairs and FlyingThings and then fine-tuned on KITTI15-training dataset. AutoFlow [14] pre-train RAFT on their proposed AutoFlow dataset and follow the same procedure to further fine-tune the model. We also pre-train the RAFT network using our RF-train dataset generated by RealFlow and then fine-tune on KITTI15-training dataset. We compare our results with them on the public online benchmark KITTI 2015 testing. The percentage of erroneous pixels (F1) is used as the evaluation metric. In Fig. 9, we show some qualitative comparison results of our method with AutoFlow and the original RAFT. As can be seen, our method achieves better results than AutoFlow and RAFT.

6.4 Visualization of RealFlow samples

Our RealFlow can automatically generate training pairs from videos without human involvement. Thus, huge amount of videos can be used to generate optical flow training pairs which help the supervised networks generalize to various scenes.

Fig. 10 shows some samples generated from ALOV [13] dataset. ALOV dataset contains diverse real-world video sequences that cover various circumstances. Fig. 11 shows samples generated from BDD100K [18] dataset. BDD100K is a large-scale diverse driving video database that covers diverse urban scenes. The samples generated from ALOV and BDD100K are used to construct our dataset RF-AB.

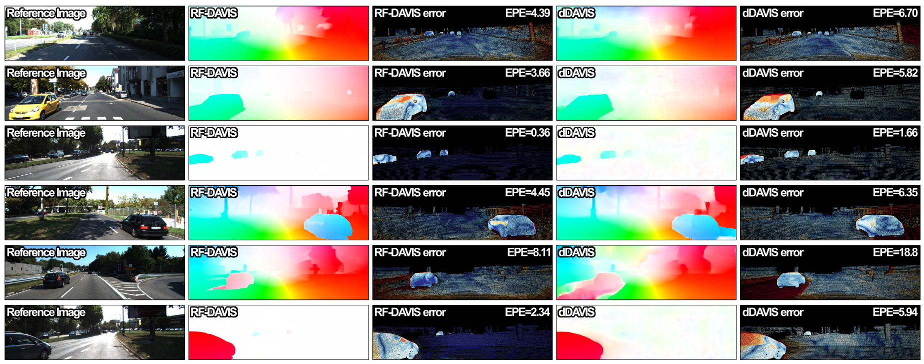
Fig. 12 shows some samples generated from DAVIS [3] challenge 2019. DAVIS consists of high-quality video sequences under various kinds of scenes. The samples generated from DAVIS are used to construct our dataset RF-DAVIS.

Besides, in order to demonstrate the efficiency and versatility of our method, we also use Vimeo-90k [16] dataset to generate some samples which is shown in Fig. 13. Vimeo-90k Dataset is a large-scale, high-quality video dataset for video enhancement such as temporal frame-interpolation, denoising, and super-resolution, which contains many close-ups. These samples may be helpful for some specific tasks.

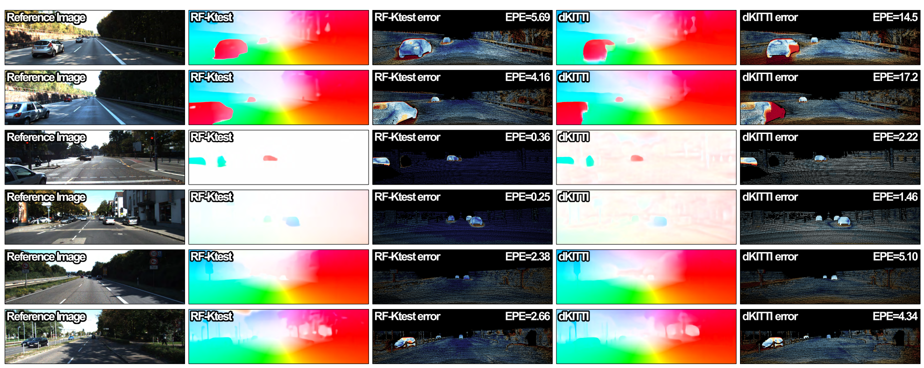
References

1. Aleotti, F., Poggi, M., Mattoccia, S.: Learning optical flow from still images. In: Proc. CVPR. pp. 15201–15211 (2021) 2, 6, 7, 11
2. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proc. ECCV. pp. 611–625 (2012) 2
3. Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Van Gool, L.: The 2019 davis challenge on vos: Unsupervised multi-object segmentation. arXiv:1905.00737 (2019) 1, 8, 16
4. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P.v.d., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proc. ICCV. pp. 2758–2766 (2015) 1
5. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proc. CVPR. pp. 4340–4349 (2016) 1
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proc. CVPR. pp. 3354–3361 (2012) 2
7. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: Proc. CVPR. pp. 7760–7768 (2020) 4, 5
8. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. In: Proc. CVPR. pp. 1–8 (2008) 5
9. Luo, K., Wang, C., Liu, S., Fan, H., Wang, J., Sun, J.: Upflow: Upsampling pyramid for unsupervised optical flow learning. In: Proc. CVPR. pp. 1045–1054 (2021) 7, 12
10. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proc. CVPR. pp. 4040–4048 (2016)
11. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proc. CVPR. pp. 3061–3070 (2015) 2
12. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proc. ICCV. pp. 12179–12188 (2021) 3, 4
13. Smeulders, A.W., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Trans. on Pattern Analysis and Machine Intelligence **36**(7), 1442–1468 (2013) 1, 8, 14
14. Sun, D., Vlasic, D., Herrmann, C., Jampani, V., Krainin, M., Chang, H., Zabih, R., Freeman, W.T., Liu, C.: Autoflow: Learning a better training set for optical flow. In: Proc. CVPR. pp. 10093–10102 (2021) 5, 6, 7, 8
15. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Proc. ECCV. pp. 402–419 (2020) 5, 8
16. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019) 9, 17

17. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proc. CVPR. pp. 899–908 (2019) 1, 6, 7
18. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proc. CVPR. pp. 2636–2645 (2020) 1, 8, 15



(a) Qualitative comparison of RAFT trained on our RF-DAVIS dataset with that trained on dDAVIS [1] dataset.



(b) Qualitative comparison of RAFT trained our RF-Ktest dataset with that trained on dKITTI [1] dataset.

Fig. 7: Qualitative comparison of our RealFlow with previous dataset generation method Depthstillation [1]. Depthstillation generated optical flow dataset dDAVIS and dKITTI from DAVIS and KITTI multiview test. We also use the DAVIS and KITTI multi-view test videos to generate our RF-DAVIS and RF-Ktest datasets. We train the same model RAFT on these datasets and evaluate on KITTI 2015 train set. Error maps are also visualized, where correct pixels are displayed in blue and wrong ones in red.

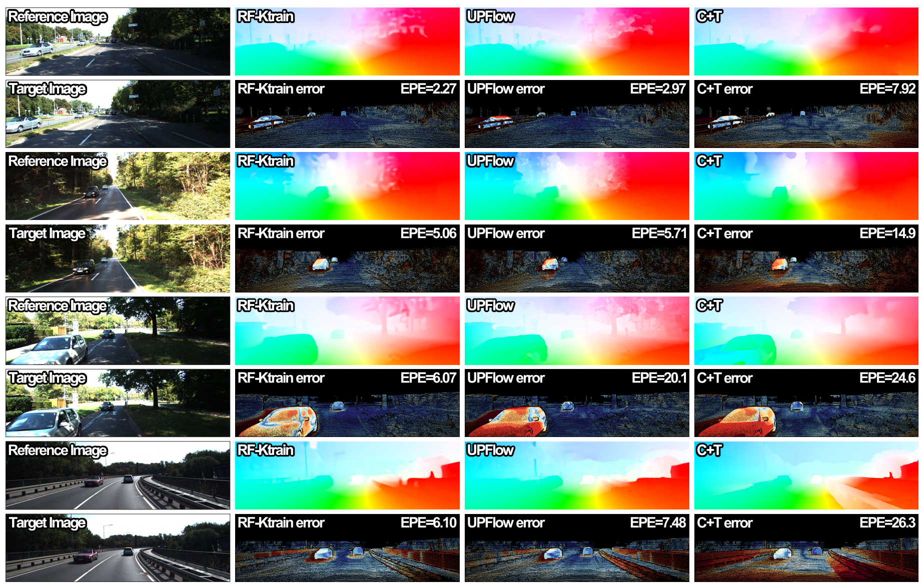


Fig. 8: Qualitative comparison of our RealFlow with unsupervised methods. We first generate RF-Ktrain dataset using the same training videos as the unsupervised method UPFlow [9]. Then we train RAFT on our RF-Ktrain to obtain the optical flow predictions for comparison with the unsupervised method UPFlow and the baseline method, where RAFT is trained on C+T (synthetic dataset FlyingChairs and FlyingThings). Error maps are visualized, where correct predictions are displayed in blue and wrong ones in red. The End-Point-Error (epe) is used as the evaluation metric, which is also depicted in the top-right side of each sample.

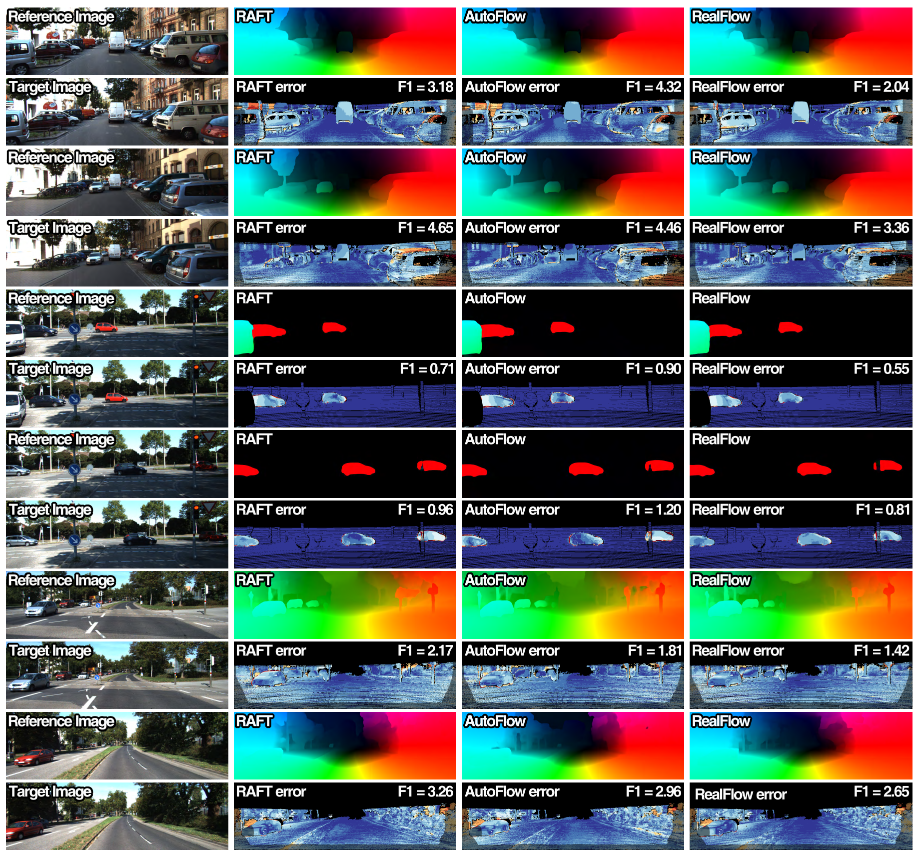


Fig. 9: Qualitative comparison of our method with AutoFlow and RAFT on KITTI 2015 online benchmark. The percentage of erroneous pixels (F1) is used as the evaluation metric. Error maps are visualized by KITTI 2015 website, where correct pixels are displayed in blue and wrong ones in red.

Fig. 10: Samples from RF-AB. These samples are generated from ALOV [13] which is a large-scale real-world video dataset that covers various circumstances. First/fourth row: image 1; second/fifth row: generated new image 2; third/sixth row: visualized optical flow. Note that, these flow maps are the ground truth of our generated pairs.

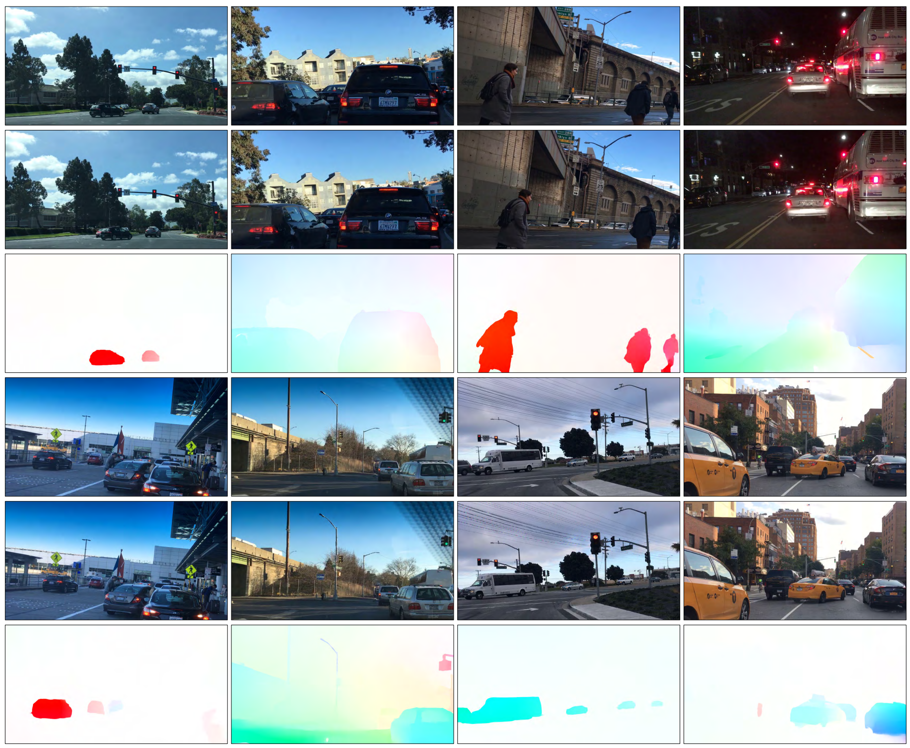


Fig. 11: Samples from RF-AB. These samples are generated from BDD100K [18] which is a large-scale diverse driving video database covering various urban scenes. First/fourth row: image 1; second/fifth row: generated new image 2; third/sixth row: visualized optical flow. Note that, these flow maps are the ground truth of our generated pairs.

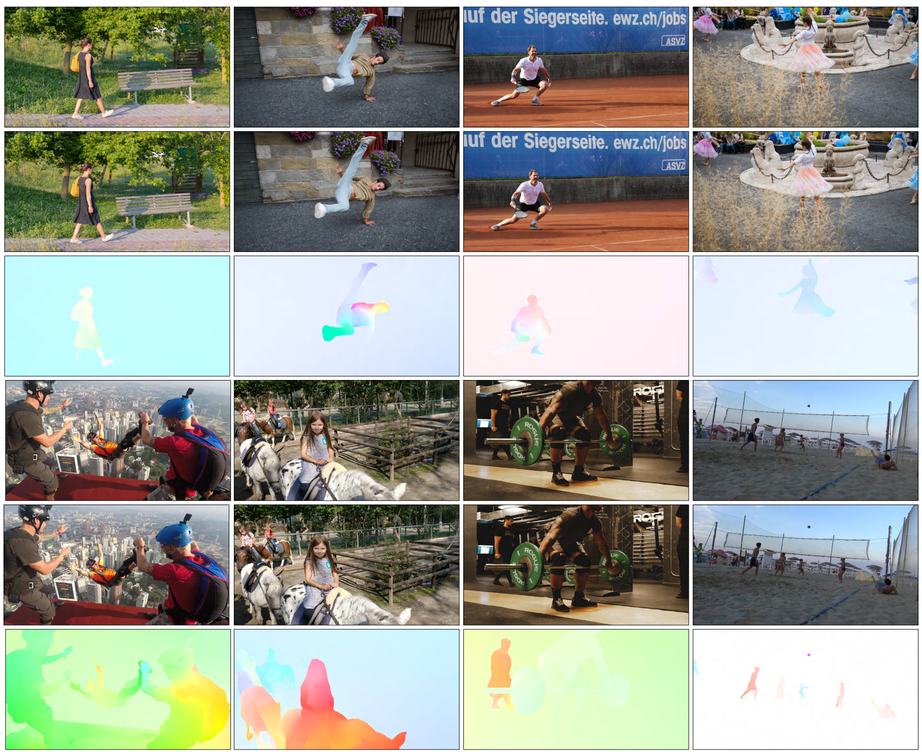


Fig. 12: Samples from RF-DAVIS. These samples are generated from DAVIS challenge 2019 [3] which consists of high-quality video sequences under various kinds of scenes. First/fourth row: image 1; second/fifth row: generated new image 2; third/sixth row: visualized optical flow. Note that, these flow maps are the ground truth of our generated pairs.

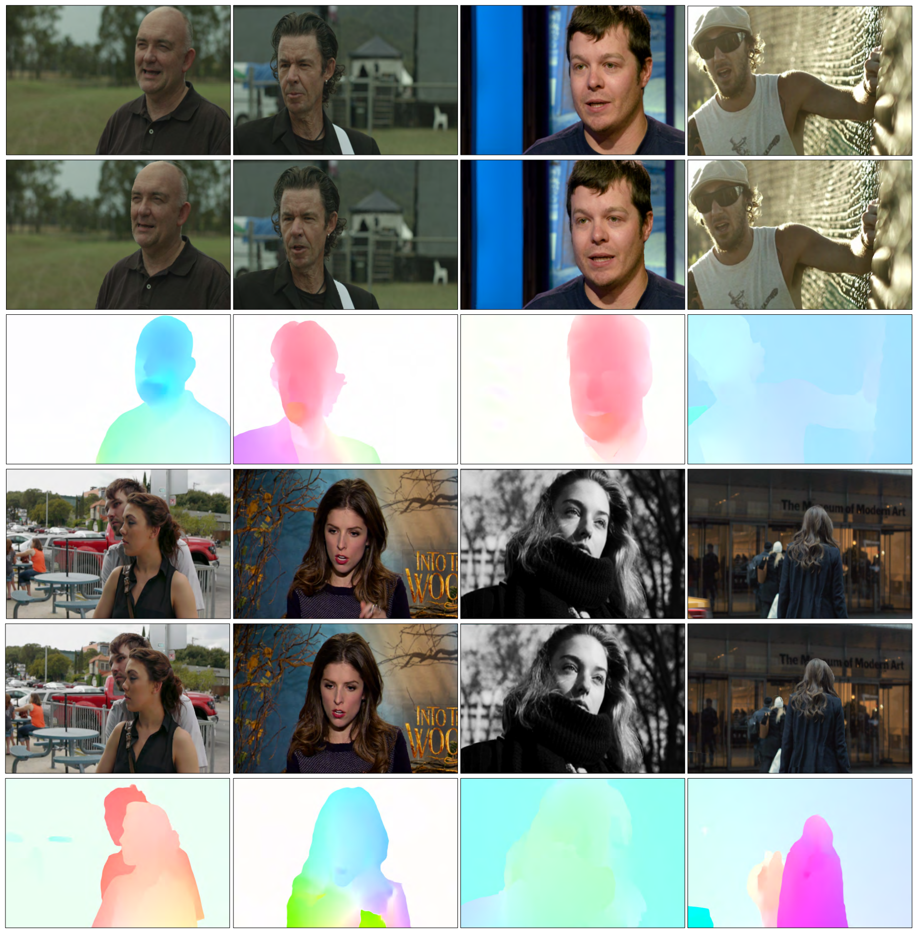


Fig. 13: Vimeo-90k Dataset [16] is a large-scale, high-quality video dataset collected for video enhancement tasks such as temporal frame-interpolation, denoising and super-resolution. We also generated some training pairs from Vimeo-90k which may be helpful for some specific scenes. First/fourth row: image 1; second/fifth row: generated new image 2; third/sixth row: visualized optical flow. Note that, these flow maps are the ground truth of our generated pairs.