RAFT: Recurrent All-Pairs Field Transforms for Optical Flow

Wei-Xin Shih
311551107

Yueh-Hsun Chiang
311553008

Qian-You Zhang
311552007

Department of Computer Science,
National Yang-Ming Chiao-Tung University

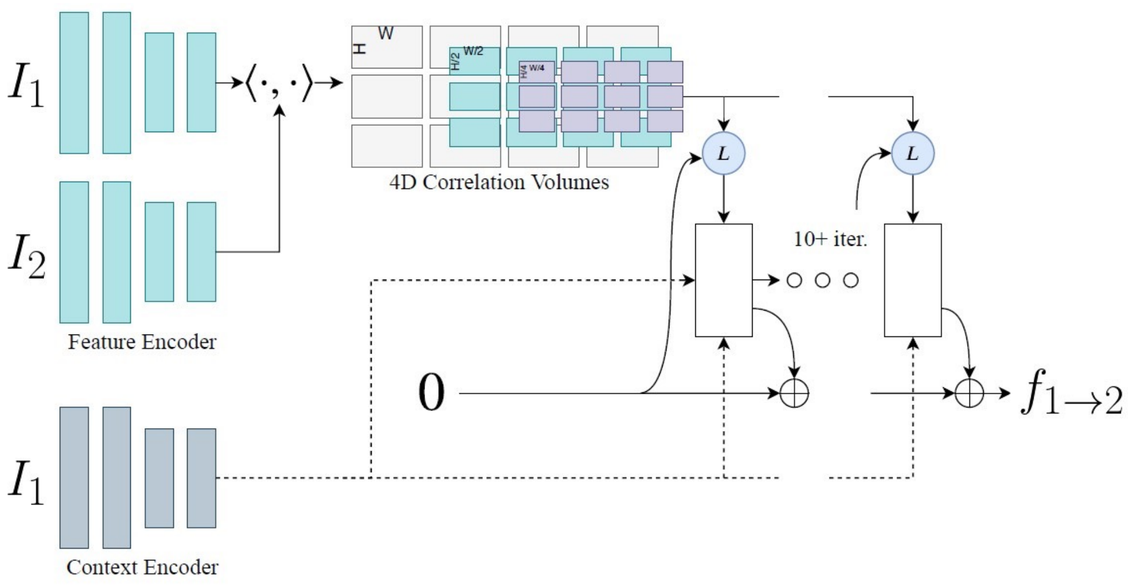
Introduction

Optical flow plays an important role in computer vision field, it enables applications such as predicting objects movement in autonomous driving, or interpolating frames to increase the frame rate of video games.

Recent researches deploys deep neural networks to estimate optical flow. However, unlike depth datasets, which ground-truth can be easily obtained by using a depth sensor like LiDAR, labeling an optical flow dataset requires a huge amount of human labor. Thus, most optical flow datasets are synthetic, and training on only these datasets will result in a domain gap between real-world scenarios.

In our experiments, we train the RAFT model in a self-supervised manner, without relying on properly annotated datasets.

Model Architecture



Methodology

The original implementation from RAFT uses L1 loss between the ground truth and the prediction series to supervise the network, with a decay factor λ:

$$L = \sum_{i=1}^{T} \lambda^{i-T} \left\| f_{gt} - f_i \right\|_1$$

In our experiments, we leverage the photometric consistency of the original image and the backward-warped image to train the network in a self-supervised manner. This enables us to train the optical flow network from scratch or fine-tune the pretrained model to give better estimate results in real-world scenario.

$$L = \sum_{i=1}^{T} \lambda^{i-T} L_{photo}$$

$$L_{photo} = \alpha \frac{1 - SSIM(I_1, \hat{I}_1)}{2} + (1 - \alpha) \left\| I_1 - \hat{I}_1 \right\|_1$$

Here $\hat{I}$ denotes backward-warped image, and we follow previous works setting $\alpha = 0.84$.

Experiments

To give a fair comparison between the performance of two different settings, we split the Sintel clean dataset, train and validate both settings using our split. 4 out of 23 scenes are extracted and used as validation set, while the others are used as the training set.

The following table reports our validation scores of the models:

	EPE	1px	3px	5px
Supervised	3.38	0.823	0.911	0.932
Self-supervised	10.82	0.698	0.798	0.822

Instead of bilinear upsampling, RAFT uses learnable convex upsampling technique, and we noticed a slight difference here between supervised and self-supervised results.

We can see from the images above, that the upsampling effect of the supervised result is visibly worse than the self-supervised one. Based on this observation, we want to know if the photometric consistency used in self-supervised manner actually boosts up the learning of upsampling module.

From this experiment, we can say that the photometric consistency can benefit the learning of the upsampling module during early phase of training.

Our code and more results
are available through these links.

Conclusion

In this final project, we reproduced the result of the widely used optical flow backbone network RAFT, in both supervised and self-supervised manners. By running in self-supervised manner, it enables us to use easily-accessible resources, such as videos recorded by smart phones to train the network. Moreover, we provide some extra observations from our experiments about the upsampling mechanics.