Paper

Title:

Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation

Authors:

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, Jan Kautz, UMass Amherst

Link:

Paper: https://arxiv.org/abs/1712.00080

Presentation on CVPR: https://www.youtube.com/watch?v=rkSzRbM4lIM

Tags:

#CNN, #slowmotion, #deeplearning, #videoframesinterpolation, #U-Net, #tagsfortags, #likesforlikes, #nofilters

Year:

2018

Code:

I have not found any code for this

Summary

What:

Generating intermediate video frames between existing ones to form spatially and temporally coherent video sequences using bi-directional optical flows and soft visibility maps created by U-Net neural networks.

Existing methods mainly generate only one intermediate frame. This paper describes a method of inserting any number of intermediate frames at any timestamps between two known frames.

This approach may be used to generate high-quality slow-motion videos from videos with any frame-rate-per-second (30fps from 24fps, 60fps from 30fps, 240fps from 30fps, 480fps from 60fps, etc).

How:

Given two input images I_o and I_1 and a time $t \in (0, 1)$ a goal is to predict the intermediate image I_t at time $T = t$.

The first step is to find matrices of graphical flow $F_{t\to0}$ from I_t to I_0 and $F_{t\to1}$ from I_t to I_1. As there is no I_t yet, those matrices are computed from $F_{0\to1}$ and $F_{1\to0}$.

This approximation works well in smooth regions but poorly around motion boundaries, because the motion near motion boundaries is not locally smooth. To reduce artifacts around motion boundaries, which may cause poor image synthesis, they propose learning to refine the initial approximation. This is done using sub-network, which produces refined $F_{t\to0}$ and $F_{t\to1}$ together with visibility maps $V_{t\leftarrow0}$ and $V_{t\leftarrow1}$. Visibility map shows from which frame the pixel is taken from.

Training was done on 1.1K video clips, consisting of 300K individual video frames with a typical resolution of 1080 × 720 taken from Adobe240-fps and YouTube240-fps datasets.

They introduced own loss

$$l = \lambda_r l_r + \lambda_p l_p + \lambda_w l_w + \lambda_s l_s$$

which consists of four different parts:

1. *Reconstruction loss* l_r models how good the reconstruction of the intermediate frame is
2. *Perceptual loss* l_p added in order to preserve details of the predictions and make interpolated frame sharper
3. *Warping loss* l_w added in order to model the quality of the computed optical flow
4. *Smoothness loss* l_s added in order to encourage neighbouring pixels to have similar flow values

The weights of lambdas have been set empirically.

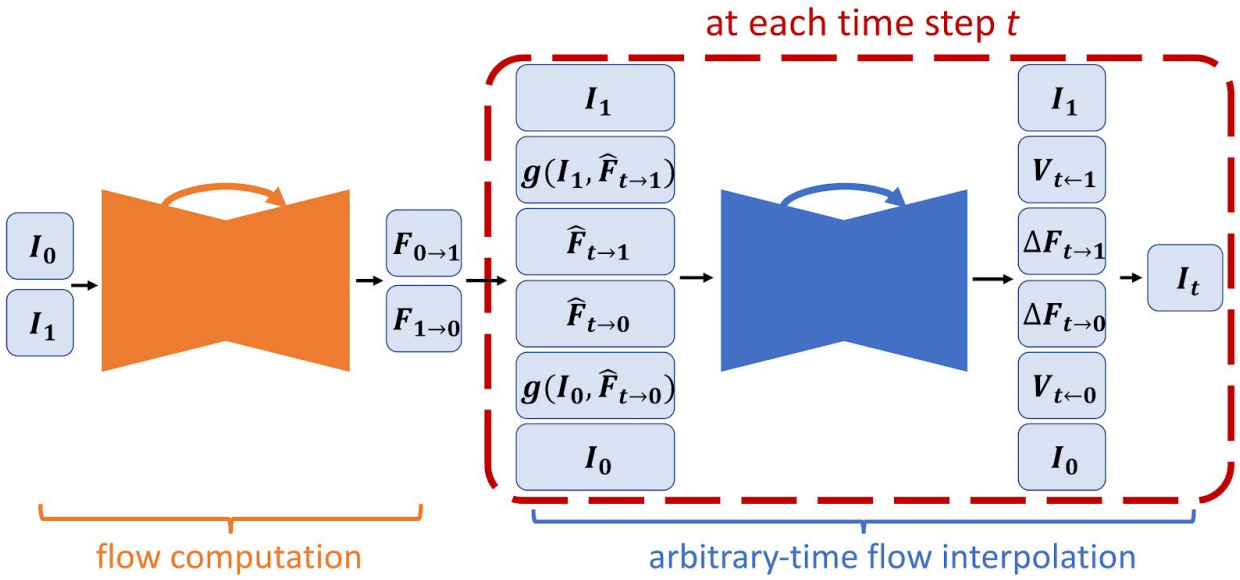

Figure 1. Network architecture

Results:

https://people.cs.umass.edu/~hzjiang//projects/superslomo/superslomo_public.mp4

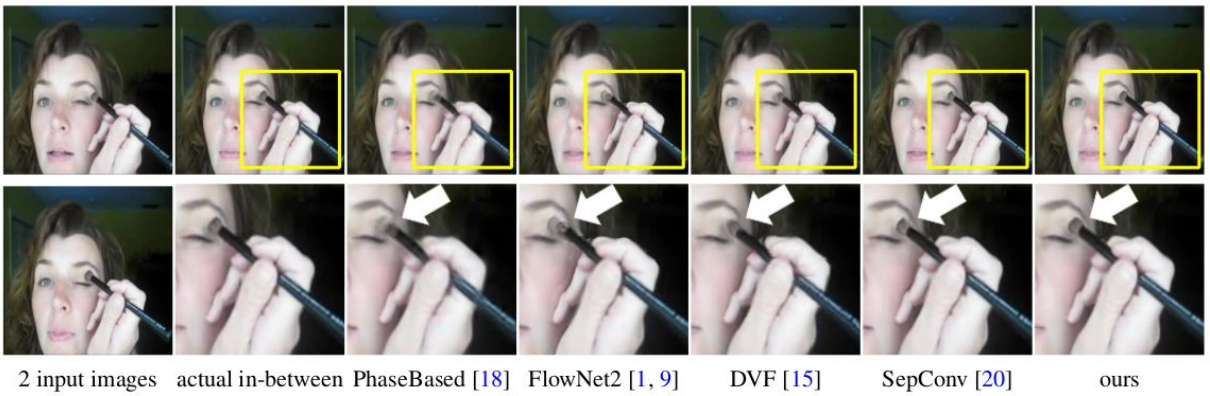

Figure 2. Visual results of a sample from UCF101 dataset compared with results of other models. Their model produces fewer artifacts around the brush and the hand.