Research Paper Summary

Md.Zarif Ul Alam
1705010

Depth Aware Video Frame Interpolation

Wenbo Bao, Wei-Sheng Lai, Chao Ma
Xiaoyun Zhang, Zhiyong Gao ,Ming-Hsuan Yang
Institute of Image Communication and Network Engineering , Shanghai Jiao Tong University
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
University of California, Merced
Google

Introduction

The idea of video frame interpolation has been prevalent for quite a long time. The problem is defined like this : from a given set of images / video , synthesize non-existent frames in-between the original frames . Before the deep learning era people tried to solve this problem using optical flow .But computing the global optical flow and interpolating the video from this information is computationally expensive and non effective in real world application . Among the previous deep learning based approach , Long et al. trained a generic CNN to directly synthesize in between frames . This however suffer from severe blurriness . Liu et al. used a 3D optical flow based approach which warps input frames based on a tri-linear sampling . This approach suffers less blurriness , but is not able to perform well in scenes with large motion . Here this paper comes to play . The paper discusses a technique to develop a depth aware flow projection layer . Using the depth information , this algorithm explicitly detects occlusion (large object motion) . The proposed algorithm of the authors is based on a simple observation that the closer objects should be preferably synthesized in the intermediate frame .

High Level Overview

As discussed in the introduction , CNN and 3D optical flow based approaches can't seem to handle large motions . To handle large motion , several other approaches use a coarse-to-fine strategy or adopt advanced deep learning based flow estimation based architecture like PWC-Net to estimate the optical flow more accurately and efficiently . A straightforward approach to handle occlusion is to use an occlusion mask for adaptively blending the pixels .However all these approaches depends on the model's capacity to implicitly detect occlusion . In this paper the authors propose an algorithm to explicitly detect occlusion by exploiting the depth information .

First a bi-directional optical flow and depth maps are generated from the two neighbouring input frames . To warp the input frames , the authors adopt a flow projection layer to generate intermediate flows . The depth information is used here . As multiple flow vectors may encounter at the same position , the contribution of each flow vector is calculated using the depth value for aggregation . This depth aware flow projection layer is the core concept of the algorithm because in contrast to simple average of flows , the proposed depth aware layer is able to generate flow with clearer motion boundaries .

Architecture and Model Overview

The proposed model has the following sub-modules :

- Flow Estimation
- Depth Estimation
- Context Extraction
- Kernel Estimation
- Frame Synthesize Network

The depth aware flow projection layer is used to obtain the intermediate flows and then the adaptive warping layer is used to warp the input frames , depth maps and contextual features . The last sub-module Frame Synthesize Network generates the output frame with residual learning .

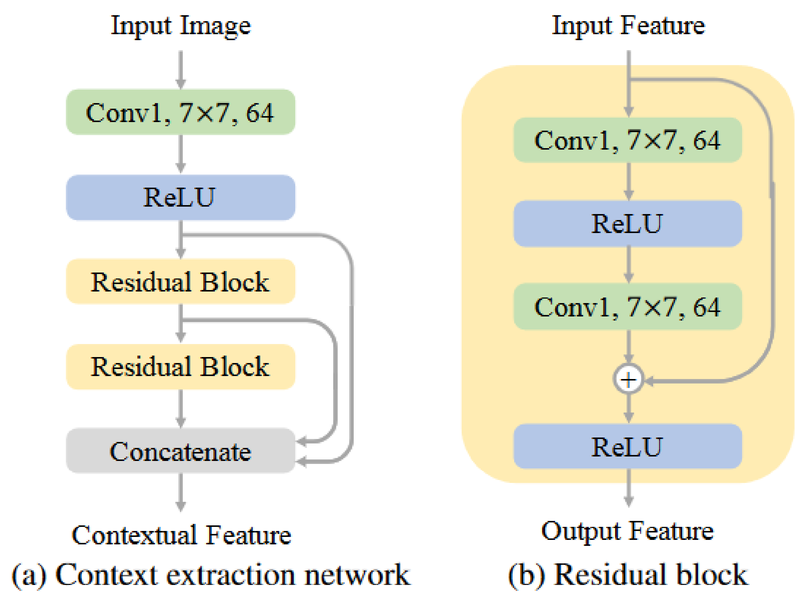

Figura 1: Structure of the Context Extraction Network

Evaluation Datasets

Middlebury : Widely used to evaluate video frame interpolation methods. Image Resolution 640 x 480 pixels

Vimeo90K : Contains 3782 triplets in the test set. Image Resolution 448 x 256 pixels

UFC101 : Contains videos with large variety of human actions . Resolution 256 x 256 pixels

Table 3. **Quantitative comparisons on the Middlebury EVALUATION set.** The numbers in red and blue represent the best and second best performance. The proposed DAIN method performs favorably against other approaches in terms of IE and NIE.

Method	Mequon		Schefflera		Urban		Teddy		Backyard		Basketball		Dumptruck		Evergreen		Average	
	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE	IE	NIE
EpicFlow [29]	3.17	0.62	3.79	0.70	4.28	1.06	6.37	1.09	11.2	1.18	6.23	1.10	8.11	1.00	8.76	1.04	6.49	0.97
SepConv-L_1 [25]	2.52	0.54	3.56	0.67	4.17	1.07	5.41	1.03	10.2	0.99	5.47	0.96	6.88	0.68	6.63	0.70	5.61	0.83
ToFlow [39]	2.54	0.55	3.70	0.72	3.43	0.92	5.05	0.96	9.84	0.97	5.34	0.98	6.88	0.72	7.14	0.90	5.49	0.84
Super SloMo [14]	2.51	0.59	3.66	0.72	**2.91**	0.74	5.05	0.98	9.56	0.94	5.37	0.96	6.69	0.60	6.73	0.69	5.31	0.78
CtxSyn [23]	**2.24**	**0.50**	**2.96**	**0.55**	4.32	1.42	**4.21**	0.87	9.59	0.95	5.22	0.94	7.02	0.68	6.66	0.67	5.28	0.82
MEMC-Net [2]	2.47	0.60	3.49	0.65	4.63	1.42	4.94	0.88	8.91	0.93	**4.70**	0.86	6.46	0.66	6.35	**0.64**	5.24	0.83
DAIN (Ours)	2.38	0.58	3.28	0.60	3.32	**0.69**	4.65	**0.86**	**7.88**	**0.87**	4.73	**0.85**	**6.36**	**0.59**	**6.25**	0.66	**4.86**	**0.71**

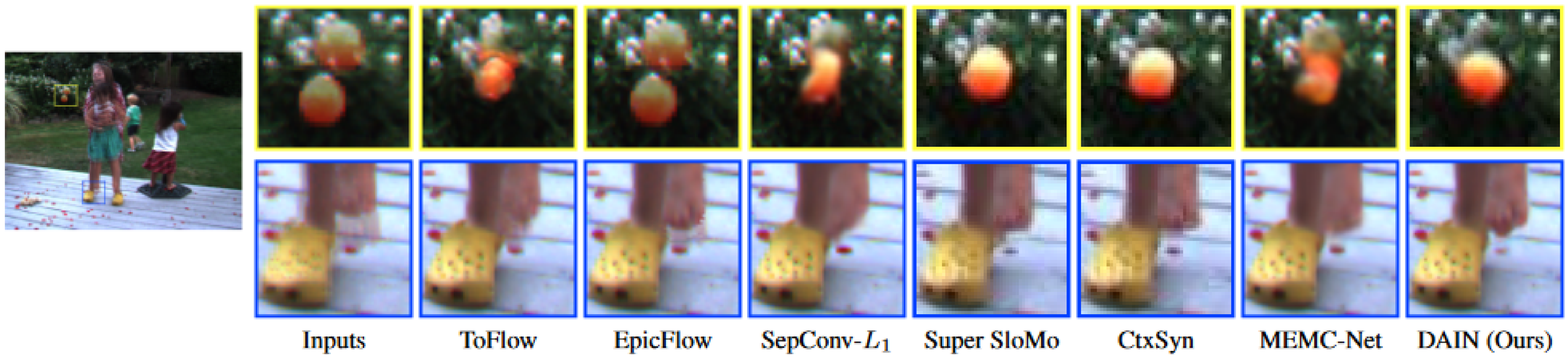

Figura 2: Comparison on the Middlebury Evaluation set

Evaluation Metric

Interpolation Error (IE) , Normalized Interpolation Error (NIE) . Lower Interpolation errors indicate higher performance .

Where does it stand compared to the SOTA models ?

The proposed model outperforms most of the current state of the art models. This model ranks 1^{st} in terms of NIE and 3^{rd} in terms of IE among all published algorithms on the Middlebury Website . A comparison is shown in Figure 2.

Real Life Applications

Video Frame Interpolation has quite a lot of real life applications and it is one of the important computer vision problems right now . Some of the use case are :

- Slow Motion Generation
- Novel View Synthesis
- Frame Rate Up Conversion
- Frame Recovery in Video Streaming

Among the use cases mentioned here , frame rate up conversion has become very popular . The higher the frame rate , the smoother the images are and they are less choppy . So , increasing the frame rate using video interpolation is a very popular use case . This frame rate up conversion is specially important in games as this gives a better gaming experience to gamers .

One other use case of this frame rate up conversion is that it can be used to make stop motion animations with higher frame rate . Stop motion animations are usually difficult to make because of the immense amount of image that an animator has to generate to make it even as low as 15 FPS . Using this video interpolation technique the huge overload of the animator can be reduced and it also makes the animation smooth and gives the viewers a better experience .

Other use case is generating slow motion video footage . It is hard to generate slow motion videos without proper camera equipment . But this video interpolation technique can leverage the depth information to generate high quality slow mo videos which can be highly effecting in real world scenarios like crash analysis .

Limitations

As discussed earlier , the algorithm uses depth map to detect occlusion for flow aggregation . However there are some corner cases , where the model doesn't seem to perform well which leads to ambiguous object boundaries and thus blurry interpolation .

Final Remark

The idea of using depth information to interpolate videos is relatively new , and the potential seems to be indicate that we will be able to get an industrially viable software 1 or 2 paper down the line . Fascinating time to live in !

Reference

Original Paper Link : `https://arxiv.org/pdf/1904.00830.pdf`