SpiritNet: Blowing spirit into image

Depth-and-flow-Aware Live Photo/Video generation

Seoul Natl Univ.
Department of Electrical &Computer Engineering
Department of Mechanical Engineering

Abstract

In computer vision, video prediction is one of the most challenging tasks due to massive amount of data and unpredictable movements. Video prediction tasks can be separated into 3 big branches: video frame interpolation, next frame prediction, and video generation. These three tasks seem quite similar, but it has a significant difference. While video frame interpolation and next frame generation based on previous video both have initial optical flow information, video generation doesn't. In our research, we propose a SpiritNet model to generate random initial flow and iterate the following procedure: prediction of next flow based on previous image and flow, and construction of next image based on flow, depth, content, and kernel. We applied a LSTM-like structure to train this sequence of data. Our main breakthroughs are 1. Applying learning method to fill "holes" after DAFI, which makes the model fast, 2. Depth-Aware Flow Randomization to generate initial optical flow, 3. Double Backpropagation, which contributes to making more natural optical flows. We were able to make plausible results from one input image to GIFs. But in some situations, we've found limitations: object collision, poor image construction when occluded image is exposed, and awkward facial expressions. We concluded that we need to apply more learning models to solve these problems

1. Introduction

New-frame prediction field, including video prediction and frame interpolation is one of major fields in computer vision. Especially, frame interpolation task has large industrial potential, especially in computer graphics or video enhancement fields. While various frame interpolating methods with remarkable results has developed and been introduced lately, video prediction still shows "awkward flow" problems [1] and generated videos looks bizarre, even if learning-based approaches are used. Also, previous video prediction methods have a limitation that high-performance comes out from using multiple images, which is not plausible situation.

We assume that these problems occur from two main reasons: one is non-sequential properties of proposed models, and one is that while flow is a crucial factor for generating plausible videos, they did not focus on it. As far as we know, all proposed video generation models generate flows in a very naïve manner or even does not consider flow at all.

We propose a novel video-generation algorithm: SpiritNet, that generates a video (serial of images) from just one image by using sequential structure, without exhibiting any "awkward flow" problem. We have done this by focusing on generating flows from previous flows. We show that learning methods should be used in flow generation carefully, along with appropriate non-learning techniques we have proposed: Depth Aware Flow Initialization (DAFI). We show how DAFI initializes flow from previous flow by using inverse depth. Also, we propose Depth Aware Flow Randomization (DAFR), which builds randomized initial flow, which makes generation of video task possible.

2. Related Work

2.1. RAFT

Recent studies of computer vision tend to cope various tasks with learning-based methods. Tasks include flow estimation, depth estimation and image reconstruction, which are tasks we need for next-frame prediction in our proposed network. In our novel network, we need to train flow-generation models strongly, and thus we need optical flow ground truth between two images. We used a pretrained model proposed by Princeton research team called RAFT: State of The Art (SOTA) flow estimation model between two frames, naively based on recurrent neural network structures. [2] Note that we did not use RAFT structure in our model, and we used it only to construct ground truth for training. In our model, we used simple CNN structure for flow reconstruction.

2.2. DAIN

DAIN (Depth Aware Video Frame Interpolation) [3] is one of frame interpolating network based on deep learning network. However, this network used depth image inside to

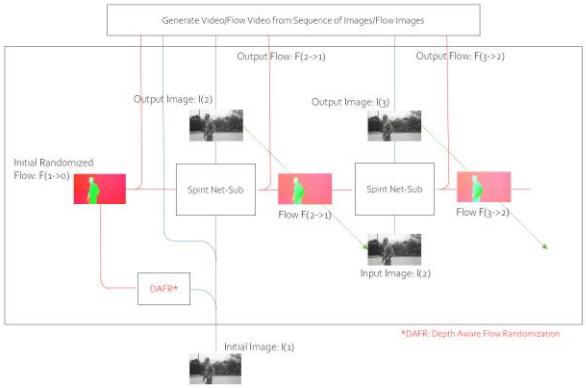
generate flow more naturally with idea using that flow of pixels should be similar when their depth is similar. So, by using correlation between depth and dynamic movements, this network successfully generated frame interpolating network. The idea we used from this model is to use depth image generating dynamic movement which is called Adaptive Warping Layer. This method is non-Learning based method which interpolates predicted image from given flow and predicted coefficients (Kernel).

2.3. Mega Depth

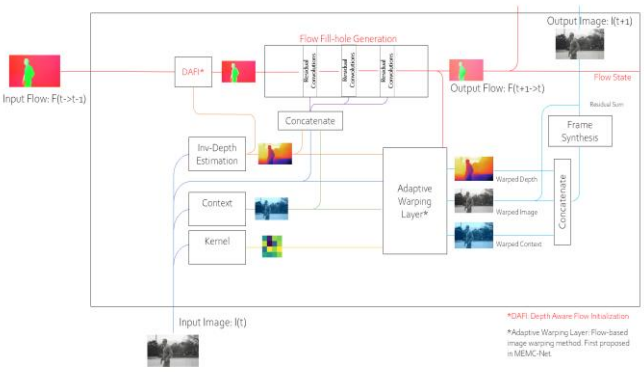
Mega Depth is SOTA Depth-Estimation model proposed by Cornell University. In our research, we used mega depth for inverse depth estimation module.

3. Spirit-Net

Our main goal is to make sequence of images seem alive by assigning plausible motion to pixels, by using only 1 image input. The procedure is mainly divided into 2 steps. First, we should generate initial flow of the image from given 1input image. This step is called DAFR (Depth Aware Flow Randomization) which we've proposed. Then, we inductively generate next image frame by input of previous image, and previous flow of its image similarly to cell state and hidden state of LSTM, we could define flow state, which passes flow of an image. We used backward flow instead of forward flow, to not make holes in warped image. You can see our model intuitively model by figure 1(a), where Spirit Net Sub containing the step mentioned earlier. By application of DAFR, we can initialize initial flow, and use image and flow as two inputs to put into submodule. From then on, we cycle through submodules with 2input&2outputs, until we get enough images/flow images to create video (flow video). For the Spirit Net Sub, as seen in fig.1(b), first step consists of extracting depth, context, and estimating kernel values for later adaptive warping. We use inverse of depth instead of depth itself since inverse depth tends to be proportional to magnitude of flow based on empirical evidence. After extracting such from current image, we first apply "Depth Aware Flow Initialization" [3]: predicting next flow in semi-naïve manner. We then put the initialized flow estimation into "Flow fill-hole generation" model. This model passes initialized flow into 3 residual convolution layers, with other features in iterative manner. We get output-flow from this. Then, we send images and feature images into Adaptive Warping Layer [3] with flow information to create warped images and feature images. We use learned interpolation coefficients to enhance performance. After that, we concatenate warped images/feature images and send it through Frame-synthesis model, to reconstruct and "smoothen out" image, to get next image. Details of specific methods are explained as below.



(a) Overview model of SpiritNet



(b) Model of SpiritNet Sub

Fig1. Overview models of SpiritNet. (a) is describing full model of spirit net and (b) takes more specific look on Spirit Net Sub Network which takes Flow and Image as a input to generate next image and flow.

3.1 Depth-Aware Flow Initialization

In frame interpolation tasks, DAIN used Depth-Aware Flow Projection method to predict(interpolate) the optical flow of intermediate image based on optical flows of neighboring frames. Unlike interpolation tasks, our goal is not to estimate flows of intermediate frames from two flows, but to predict the next flow from only one flow (which is the previous flow). So, to predict the next flow, we apply some naïve assumptions that flow would be consistent from the previous one and only the pixels would be parallel shifted backwards in direction of each optical flow. We could easily predict the next flow with this naïve method, but some problems may occur. If more than two of the shifted flows are assigned to one pixel, we cannot simply choose one of the flows. Also, since the number of pixels and their corresponding optical flows are consistent, this flow collision issue directly leads to the fact that none of the flows are assigned to some pixels, which means holes exist in the predicted flow image. To handle those two

issues, we apply our new method: Depth-Aware Flow Initialization and learning-based filling holes.
First, to handle the flow collision issue, we do not choose one of the flows assigned, but use weighted sum of the flows. Based on idea that we could observe relatively shallow parts are more dynamic while relatively deep parts are more static, we use reciprocals of depth as the weight. This method can be represented in mathematical expression as equation1 below.

$$F_{t+1\to t} = \frac{\sum_{y\in S(x)} \frac{F_{t\to t-1}(y)}{D(y)}}{\sum_{y\in S(x)} \frac{1}{D(y)}}$$

Eqn.1 S(x) is a set of warped pixels where x is a pixel of current image. For pixels y of next image where more than two flows collide, the equation calculates weighted average using reciprocal of depth 1/D(y) as weights.

For handling the holes, previous studies applied outside-in strategy, which is near neighbor likely method. But in practice, this method is slow and inaccurate since it needs scanning for all the pixels in flow image. So, we used learning model to fill the holes. This so-called Flow Fill-hole generation model uses residual convolution layers (similar to ResNet structure) with iterative image, context, and reciprocal of depth update to learn next flow better. In summary, we initialized predicted flow using DAFI and applied Flow Fill-hole generation, which is a learning model.

3.2 Depth-Aware Flow Randomization

Applying the previous method, we can predict next flow from previous flow and image. Since we need two neighboring images to get the optical flow and our goal is to construct images from just one image, we need to build initial flow information for next image construction. Applying similar idea from previous DAFI method, we use distribution of inverse depth to construct the initial flow. After making histogram of inverse depth distribution, we separate this histogram with clusters applying some thresholds. In our research, we used top-k difference: which means sorting distribution and cutting the histogram into intervals with boundaries of big difference. (But this threshold also could be learned for improvement.) For pixels in same cluster, which means they are in same group of reciprocals of depth, we assign low-variance value proportional to sigmoid of reciprocal of depth to magnitude of flow, and low-variance angle to orientation of flow for randomization, based on similar idea of DAFI. This way, we could build a plausible initial flow.

Representing DAFR in simple way looks like below:

$$|flow| = \sigma\left(\frac{1}{d}\right) \times f_0 + \epsilon$$

$$\angle flow = e^{f\left(\frac{1}{d}\right)} \times \theta_0 + random(0,2\pi) + \epsilon$$

Where functions f, and σ are custom functions you can choose. Functions f and σ must satisfy following properties:

1. f, and σ needs to “cluster” when reciprocals of depth (input) is close enough values
2. f, and σ needs to be distinguished enough when reciprocals of depth (input) are in different intervals of histogram inverse-depth map

3.3 Double Backpropagating

For previous research of image generation tasks, most of them only uses loss of images. It is a quite intuitive idea since our task is to build sequence of images similar to real videos, but we’ve observed the importance of optical flow. Actually, our model only generates estimated sequence of optical flows, not the whole image. For construction of next images, we apply adaptive warping layer which uses input as Depth, Context, Kernel of current image and the generated optical flow. So, we can know that flow estimation is the most significant part (almost whole) of our model. To improve performance of our model, we used loss of flows within loss of images: we used “Double Backpropagation”.

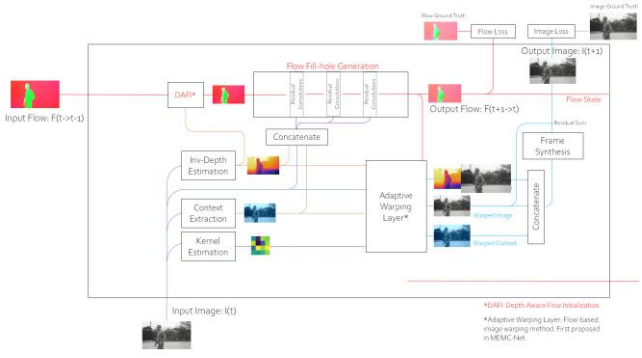


Fig2. Double Backpropagation of SpiritNet. In SpiritNet, we double backpropagates using both image loss and flow loss.

For loss of flow, we used L1-Loss (typical loss for image construction) and for loss of image, we used sum of Perceptual Style Loss and Total Variation Loss. Perceptual Style Loss is loss between features of image we've generated and image of ground truth, which contributes to making video close to ground truth (making plausible video) by reducing difference of features. We used pretrained VGG-19 for feature extraction. Total Variation Loss is sum of variation using value of every pixel of each image, which enforces smoothing penalty.

4. Experimental Results

4.1. Result

We used Vimeo 90K dataset for training, which is as open-source dataset provided by MIT, containing over 90,000 videos consisting of 7 image frames (so-called GIFs). We randomly extracted 5000 videos for training, 600 videos for validation and used 40 initial images to create videos for final testing. We trained for total 20 epochs with typical Adam optimizer with learning rate of 10^{-4}, learning rate of 10^{-7} for pretrained model Mega Depth. We successfully obtained GIFs from only one initial image and some plausible results are shown in figure 3 below. Fig 3. (a) are example input images, and (b) are corresponding output images from GIFs (6 generated frames) that SPIRIT NET has generated. Although some parts are blurred, most of the movement seems natural and more developed from previous studies.

Fig 3. (a) Example Input Images

Fig 3. (b) Example Output GIFs

4.2. Limitations

Although many results were plausible, there were some limitations found from results. First, when background moves too fast comparing to an object (or a person), background pixels affect low-depth pixels, hence objects tend to be "absorbed" to background as you can find in Fig 4. (a). This problem could be solved by improving DAFR method such as adding learnable thresholds. Next problem is when pre-occluded image becomes exposed, result of reconstruction becomes poor since there is no information of the occluded object. Last problem occurs when facial expression is in the image, the results seem non-human or awkward. This 'awkward facial expression' issue is a chronic problem in image generating tasks of computer vision. Both second and third problems could be solved by

using generative models (e.g., Generative Adversarial Networks, GAN) instead of reconstruction model (e.g., Variational Auto Encoder, VAE). Examples of limitations found if results of SpiritNet are shown below in Fig 4.

Fig 4. (a) Objects collision(absorb)

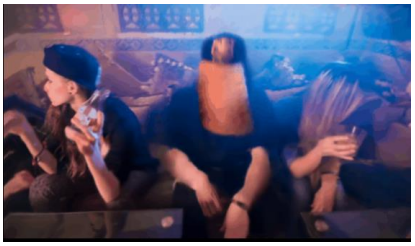


Fig 4. (b) When occluded object expose

Fig 4(c) Facial expression

Fig4. Examples of Limitations found in results of SpiritNet

4.3. Further plans

We plan to consider "semantics" in future work: by adding semantics state, which would be a vector passing through sequential structure containing semantic information, e.g. "move right". We expect adding this state would enhance performance.

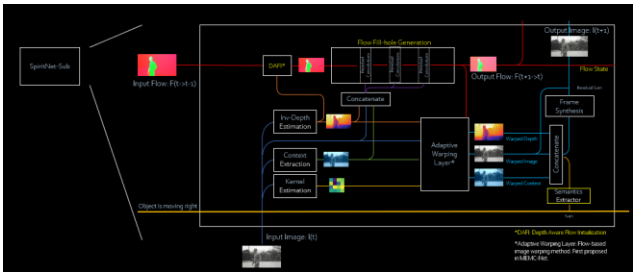


Fig5. Expected structure of enhanced SpiritNet: with semantics state added

We expect adding semantics state could expand the usage of SpiritNet: for example, we can expect SpiritNet to create video from text: by inputting text into semantics-extracting NLP model, and then using those semantics information as initial semantics state.

5. Conclusion

SpiritNet succeeded in generating GIF images from only one input image. Our model focused on constructing natural optical flows by using groundbreaking methods: DAFR, DAFI and Double Backpropagation. DAFR and DAFI used depth image obtained from Mega Depth to generate more natural movement, based on idea that dynamism is usually proportional to inverse of depth. To reinforce construction of natural flows, we added backpropagation of flow loss, which is named as Double Backpropagation. We were able to get GIFs with natural movement. However, there were some limitations found in results: object collision, poor image construction when occluded image is exposed, and awkward facial expressions. In further research, these problems may be solved by adding more learning-based method inside such as learnable thresholds is DAFR or application of GAN for natural flow construction in SpiritNet.

References

[1] Generating videos with scene dynamics. Carl Vondrick, Hamed Pirsivash, Antonio Torralba. MIT.

[2] RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Zachary Teed and Jia Deng. Princeton University

[3] Depth-Aware Video Frame Interpolation Wenbo Bao1 Wei-Sheng Lai3 Chao Ma2 Xiaoyun Zhang1∗ Zhiyong Gao1 Ming-Hsuan Yang3,4 1 Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

[4] MegaDepth: Learning Single-View Depth Prediction from Internet Photos Zhengqi Li Noah Snavely Department of Computer Science & Cornell Tech, Cornell University