A Comprehensive Study on Novel Video Frame Interpolation Methods

Hrishikesh Mahajan
Student
School of Computer Engineering and Technology,
MIT World Peace University
Pune, India

Yash Shekhadar
Student
School of Computer Engineering and Technology,
MIT World Peace University
Pune, India

Shebin Silvister
Student
School of Computer Engineering and Technology,
MIT World Peace University
Pune, India

Dheeraj Komandur
Student
School of Computer Engineering and Technology,
MIT World Peace University
Pune, India

Nitin Pise
Professor
School of Computer Engineering and Technology,
MIT World Peace University
Pune, India

ABSTRACT

Video Frame Interpolation is the process of generating frames between two or more frames of a video. This process helps in the generation of slow-motion videos or helps in increasing the framerate of the video. Today, methods such as Optical Flow, Depth mapping and Visibility Mapping techniques are used to interpolate frames of high quality with less emphasis on Learning-Based methods. Thissurvey demonstrates a comprehensive overview of major research contributions in this domain. This paper provides an overview of 18 research papers along with novel architectures. The papers are compared with respect to two benchmark datasets: UCF 101 and Vimeo 90k across two metrics: Peak signal-to-noise ratio(PSNR) and Structural Similarity Index(SSIM).

Keywords

Video Frame Interpolation, Deep Learning, Optical Flow, Video Processing

1. INTRODUCTION

Frame interpolation is the process of generating new frames between two existing frames such that the generated frames will accurately model the motion between the two pre-existing frames. There have been many conventional algorithms in existence for the task of frame interpolation, one of the most common ones being Linear Frame Interpolation (LFI), wherein the interpolated image will have the pixels in the mean position of the previous and next image. This method does not take into account the motion of objects, resulting in moving objects having multiple edges. The current popular method used in television sets is the Motion Compensated Frame Interpolation method which takes into account pixel velocity and accordingly interpolates frames to give out good approximation frames. This technique also is inadequate and generally leads to something called the “soap-opera effect”. The interpolation techniques can roughly be classified into the following four categories namely pixel-based, shape-based, registration-based and learning-based.

2. EXISTING METHODS

A. Depth-Aware Video Frame Interpolation:

Bao et al. [1] make active use of depth information to tackle the occlusion problem. This approach works very well when the objects in the image are close. The limitation of this approach is that it is making use of estimated depth maps and not true depth maps and also the approach is compute-intensive.

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

Jiang et al. [2] in their work have created a method that can create any number of frames in between two frames making the use of bi-directional optical flow and soft visibility maps. In their paper, they propose the fusing of the two input images i.e. the image at t_1 and the image at t_3 and giving both the images the appropriate amount of “weight” while fusing them together using visibility maps and optical flow calculations. This approach is compute-intensive and the training phase may change according to video requirements.

C. RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation:

Huang et al. [3] in their paper overcame the drawback of bi-directional optical flow calculation by proposing a novel model called Intermediate Flow Network (or IFNet). The IFNet is a neural network that can estimate intermediate flows between two images. A supervision or teacher model is used to boost the convergence process of IFNet. Due to the ability of the network to estimate the intermediate flows this method achieves good results in less time. This approach, however, requires us to generate frames serially and does not allow us to perform this task parallelly.

D. Video Frame Interpolation via Adaptive Separable Convolution:

Niklaus et al. [4] in their approach use a Convolutional Neural Network which accepts two input frames and estimates pairs of 1D kernels for all pixels simultaneously as opposed to other approaches which use 2D kernels leading to huge data demands. This helps in a significant reduction in storage memory. The model has been trained using the perceptual loss function. However, the generated images have relatively lower resolution and the overall model is quite complex.

E. Efficient Video Frame Interpolation Using Generative Adversarial Networks:

One of the salient features of Generative Adversarial Networks is their ability to generate photorealistic images. GANs can work very well for tasks such as image generation, style transfer, text to image generation and many such seemingly complex tasks. Tran et al. [5] make use of generative adversarial networks to generate the interpolated frame. This approach works very well and takes around ⅕ththe time as compared to conventional interpolation methods.

F. PhaseNet for Video Frame Interpolation:

The paper [6] proposes a deep learning-based approach for intermediate frame interpolation. The method combines the phase-based approach along with a learning framework using neural networks. The purely phase-based approach as used in [7], generates the intermediate frames by phase decomposition of the input images which involves separation of the amplitude and phase of each band in the image followed by temporal filtering, phase denoising and amplification or attenuation of its amplitude. This purely phased based approach is more stable to lighting conditions as compared to the optical flow-based methods but fails to handle cases when there is a wide range of motion. This drawback has been overcome using the PhaseNet method. PhaseNet takes the phase decomposition of the images as input and estimates the phase and the amplitude values of the intermediate image level by level from which the final image is reconstructed after applying the reconstruction equation. It processes the channels of the input image independently and requires a relatively small no. of parameters.

The model employs a decoder only network having two convolutional layers of kernel size 1x1 and 3x3 respectively followed by batch normalization and ReLU nonlinearity. Furthermore, the loss function used consists of two terms. First is image loss which is the L1 norm of the pixel differences. The second is phase loss which is based on the deviations in the predicted phase from the ground truth phase. The introduction of phase loss improves the image sharpness and also ensures a considerable reduction in the training time. The curriculum learning method [8] has been used while training the network model. The resultant model is robust against lighting conditions, can handle a larger range of motion and produces smooth, plausible results. However, the model does not achieve the same level of details as the methods which match and warp the pixels. The SSIM score is lower than other interpolation methods like SepConv [4] in the case of high-frequency content, but there is no perceptual difference in the generated output.

The model, having 460k parameters, was trained on Nvidia Titan X (Pascal), taking approximately 20 hours of training time on the DAVIS Video Dataset [9].

G. Context aware method:

The paper [10] uses a context-aware synthesis approach, wherein the model takes in not only the input frames but also the contextual information. A pre-trained neural network - ResNet-18 [11] is used to extract this from the input frames. Further, an optical flow algorithm - PWCNet [12] estimates the bidirectional optical flow between the frames which are later pre warped with the input frames and the context maps. The interpolated frame is generated using a context aware video frame synthesis neural network. The model uses a GridNet architecture which is well suited for pixel-wise problems. Extracting and feeding the contextual information helps in avoiding the loss of information during frame generation.

The method is capable even in cases having occlusion and rapid motion. The model uses two terms in its loss function. L1 loss calculates the difference between the predicted frame and actual frame. Second is the feature-based loss also called perceptual loss, obtained from VGG19 [13]. The dataset used is high-quality YouTube videos with a resolution of 300 x 300 pixels. The training was done on Nvidia Titan X (Pascal) and took two days to train on the dataset having 50,000 samples.

H. IM-Net for High-Resolution Video Frame Interpolation:

The paper [14] uses a fully convolutional neural network for estimating the interpolated motion vector field (IMVF) and occlusion map which extracts the features from the input frames. This paper uses a unique approach bypassing three pairs of frames at different scales, thus following a three-level pyramidal input structure. The features from the CNN are passed to an encoder-decoder architecture which synthesizes the interpolated frame at each scale, which is further merged to generate the merged output. These merged outputs are passed through three parallel estimators consisting of convolutional layers to create the occlusion map. The dataset used is the YouTube videos having HD and FHD resolution which include clips of sports events like basketball, soccer, etc. IM-NET was tested on the Vimeo dataset [15] and obtained an average PSNR of 32.35 decibels. However, in the case of higher resolution frames, it performs better than other methods like SepConv [4] and TOFlow [16]. Also, the architecture is lightweight, which helps faster processing and outperforms other state of the art methods by a factor of 16 in terms of speed.

I. Channel Attention Is All You Need for Video Frame Interpolation:

The paper [17] aims to replace the pre-existing technique of optical flow estimation. Instead, it employs a feature reshaping operation, known as PixelShuffle (with channel attention) which replaces the optical flow estimation module. The main idea behind this is to distribute the information contained in a feature map into multiple channels and extract information by attending the channels for pixel-level frame synthesis. The use of the Attention mechanism makes the neural network focus on important regions of its feature representations. The said algorithm performs equally well in presence of challenging motion and occlusion. The proposed technique outperforms the existing models that use optical flow estimation on almost all standard datasets like UCF101, Vimeo 90k and Middlebury.

J. Generating Realistic Videos from Keyframes with Concatenated GANs:

The paper [18] proposes a novel approach to generate a series of intermediate frames from two frames as input (X_0 and X_{n+1}). This method uses two concatenated generative adversarial networks in order to achieve multiple frame interpolations in one shot. The first Generator G_1 is a simple convolution-deconvolution generator and the second generator G_2 is based on the state-of-the-art U-Net architecture. The first GAN learns the motion using the normal autoencoder pipeline and the second

GAN learns frame details with the help of U-Net. The architecture employs only 2D convolution as 3D convolutions are computationally expensive. Adversarial loss, gradient difference and normalized product correlation loss are used while training the model. The experiments are performed on UCF101, Google Push and KTH dataset to validate the effectiveness of the proposed model.The proposed cat-GAN architecture can successfully generate realistic video from input frames which is supported by evaluating the results using metrics like PSNR and SSIM.

K. Video Frame Synthesis using Deep Voxel Flow:

The method proposed in the paper [19] combines the benefits of methods like pixel approximation and optical flow estimation by training a deep neural network that learns to compute intermediate video frames by flowing pixel values from existing input pixel values, which it refers to as deep voxel flow. This is a self-supervised approach and learns to reconstruct a frame by using the voxels from nearby frames, consequently more sharp and realistic results are obtained. Trilinear interpolation is performed on the input frames (t_1 and t_3) to compute the desired intermediate frame(t_2) . In this way both previous and next frames' pixel values are taken into consideration for performing frame interpolation. The architecture used is a fully convolutional encoder-decoder architecture with three convolution layers, three deconvolution layers and a bottleneck layer. Deep voxel flow is trained from the UCF-101 training set and evaluated on the KITTI odometry dataset. The algorithm is able to achieve higher values of PSNR and SSIM on both datasets. This method shows improvement on both optical flow and recent CNN techniques for interpolation

L. Long-Term Video Interpolation with Bidirectional Predictive Network:

This paper [20] focuses on solving the challenges faced in long term frame interpolation in videos. They attempt to generate multiple frames in between two input frames. This approach is unlike most existing methods which only predict one intermediate frame or specific number of frames between two given time stamped adjacent frames. The paper explains a novel architecture based on deep learning called bi-directional predictive network (BiPN). The BiPN is bi-directional in nature and interpolates frames from two opposite directions. This bi-directional approach helps the model to predict longer video frame sequences and also helps in improving accuracy. The BiPN model can also be used to generate multiple probabilistic procedures by sampling across different noise vectors. They have used a custom joint loss to train their model which takes both features' spaces and adversarial loss into account. The paper also discusses the advantages of their model by evaluating on two benchmark datasets: 2D Shapes and UCF101 which show competitive results to other published methodologies.

M. Frame Interpolation with Multi-Scale Deep Loss Functions and Generative Adversarial Networks:

The paper [21] presents a multi-scale GAN (Generative Adversarial Networks) for intermediate frame interpolation in videos. They have proposed a novel multi-scale residual network called FIGAN, which predicts flow and synthesises the frame in a coarse-to-fine fashion, this improves the efficiency of the model. The model proposes a perceptual based loss function and it consists of two content loss functions and one adversarial loss. This improves the quality of synthesised intermediate video frames. The model is evaluated on a test dataset of YouTube -8m dataset which has a frame rate of 60. The results shared by the paper show a greater accuracy than state-of-the-art models and have subjective visual quality at par with these models. The FIGAN network has a fast run time which is ×47 faster than the state-of-the-art model compared in the paper.

N. Deep frame interpolation for video compression:

This paper explains the problem faced in interpolation methods that use deep learning based optical flows to predict the intermediate frames. These methods have limitations for predicting on a dataset with complex non-translational motions and also limited to block-based motion vectors. This paper [22] proposes a deep learning based frame interpolation network that is used for video compression and it also aims to solve the previous limitations, by adjusting with different types of geometrical deformations by compensating dense motion. The experiments in the paper comparison with the classical bi-directional hierarchical video coding structure showcase the efficiency of the novel architecture over the previously known tools of the HEVC codec.

O. Video frame interpolation via optical flow estimation with image inpainting:

This paper [23] approaches the problem of image interpolation using a three-part process. In the first part, the algorithm combines the Horn-Schunck (Global) algorithm and the Lucas Kanade (Local) algorithm for obtaining the optical flow. In the second stage, the Image Warping method is used to solve the problem of large offsets. The pixel mapping function used is an inverse distance weighted interpolation algorithm. In the third and final stage, the paper proposes image inpainting wherein the algorithm will fill in the overlaps and holes generated by the optical flow algorithms by fetching data from existing frames. This paper compares this method with multiple other interpolation methods and has inferred that this method works better in terms of edge preservation, noise and artefact preservation.

P. Deep Video Frame Interpolation using Cyclic Frame Generation:

In this method [24] the interpolation is done in two stages where in the first stage is a pre-trained baseline model which accepts two edges only images and two actual images (1st and 3rd) to generate the intermediate image. In the second stage, this same network architecture is used three times with I0, I1 for the first network which will give out I (0.5), I1, I2 for the second network which will give out I (1.5). Now both I (0.5) and I(1.5) will be used for the third and final network to obtain an interpolated frame. The models are based on the Deep Voxel Flow model. Additionally, the network also incorporates two losses namely cycle consistency loss and motion linearity loss.

Q. Lap-Based Video Frame Interpolation:

In this paper [25] the method of Local All-Pass was used to calculate the optical flow between two images. LAP makes use of quadratic functions to estimate the optical flow which is what makes it so accurate as compared to other methods which only use first-order equations. It is also vastly compute efficient as compared to a Convolutional Neural Network, completing the task in 1/300th of its operations. The method also makes use of perceptive metrics rather than the MSE and SSIM which the author of the paper argues are inherently flawed when it comes to comparing interpolation techniques.

R. SoftMax Splatting for Video Frame Interpolation:

In this paper [26] the method of forwarding warping takes prominence, a technique that is often avoided because it maps many pixels to one pixel in the future which is undesirable to the task of frame interpolation. However, with the help of SoftMax Splatting this changes, as it counters the problem of assigning multiple pixels to a single pixel in the future predicted frame. Furthermore, the paper also proves that feature pyramids can be used for high-quality image synthesis. The use of SoftMax splatting facilitates the use of end-to-end training and thus assiststhe feature pyramid extractor to accumulate features that are crucial for image synthesis.

3. RESULTS AND COMPARISON

Table 1. represents the quantitative comparison of the metric scores of the models on the benchmark UCF101 and Vimeo 90k dataset.

Table 1. Metric Scores on UCF101 Dataset

Paper	**UCF 101**	
	PSNR	**SSIM**
A. Depth Aware Video Frame Interpolation [1]	34.99	0.9683
B. Super SloMo: High-Quality Estimation of Multiple Intermediate Frames for Video Interpolation [2]	33.14	0.938
C. RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation [3]	35.29	0.969
E. Efficient Video Frame Interpolation using Generative Adversarial Networks [5]	29.22	0.835
I. Channel Attention Is All You Need For Video Frame Interpolation [17]	34.91	0.969
J. Generating Realistic Videos From Keyframes With Concatenated GANs [18]	~32	~0.94
K. Video Frame Synthesis Using Deep Voxel Flow [19]	35.8	0.96
L. Long-Term Video Interpolation with Bidirectional Predictive Network [20]	31.4	0.94
P. Deep Video Frame Interpolation Using Cyclic Frame Generation [24]	36.96	0.953
R. SoftMax Splatting for Video Frame Interpolation [26]	36.1	0.97

Table 2. Metrics Scores Vimeo 90k dataset

Paper	**Vimeo 90k**	
	PSNR	**SSIM**
A. Depth Aware Video Frame Interpolation [1]	34.71	0.9756
C. RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation [3]	36.10	0.980
G. Context-Aware Synthesis for Video Frame Synthesis [10]	33.50	0.9473
H. IM-Net for High-Resolution Video Frame Interpolation [14]	33.50	0.9473
I. Channel Attention Is All You Need for Video Frame Interpolation [17]	34.65	0.973

4. CONCLUSION

This paper reviews the various techniques used for video frame interpolation. There are various mathematical and deep learning-based approaches for solving this problem. The models have been analysed and compared. Mathematical models are faster in terms of execution speed; however, deep learning-based models are more robust to conditions like rapid movement, occlusion, etc and also the predicted frame is more accurate, with fewer artefacts.

5. REFERENCES

[1] WenboBao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3703-3712

[2] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, Jan Kautz; Super SloMo: High-Quality Estimation of Multiple Intermediate Frames for Video Interpolation, CVPR 2018.

[3] Huang, Z., Zhang, T., Heng, W., Shi, B., & Zhou, S. (2020). RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation. ArXiv, abs/2011.06294.

[4] Niklaus, Simon & Mai, Long & Liu, Feng. (2017). Video Frame Interpolation via Adaptive Separable Convolution. 261-270. 10.1109/ICCV.2017.37.

[5] Tran QN, Yang S-H. Efficient Video Frame Interpolation Using Generative Adversarial Networks. Applied Sciences. 2020; 10(18):6245. https://doi.org/10.3390/app10186245

[6] Meyer, Simone &Djelouah, Abdelaziz&Mcwilliams, Brian &Sorkine-Hornung, Alexander & Gross, Markus & Schroers, Christopher. (2018). PhaseNet for Video Frame Interpolation. 498-507. 10.1109/CVPR.2018.00059.

[7] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. SorkineHornung. Phase-based frame interpolation for video. In Computer Vision and Pattern Recognition, pages 1410– 1418, 2015

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48. ACM, 2009

[9] Davischallenge.org. 2021. DAVIS: Densely Annotated Video Segmentation. [online] Available at: https://davischallenge.org.

[10] Niklaus, Simon & Liu, Feng. (2018). Context-aware Synthesis for Video Frame Interpolation.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, Dec 2015.

[12] D. Sun, X. Yang, M. Liu and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934-8943, DOI: 10.1109/CVPR.2018.00931.

[13] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, September 2014.

[14] T. Peleg, P. Szekely, D. Sabo and O. Sendik, "IM-Net for High-Resolution Video Frame Interpolation," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2393-2402, DOI: 10.1109/CVPR.2019.00250.

[15] Vimeo.com. 2021. Vimeo | The world's only all-in-one video solution. [online] Available at: https://vimeo.com.

[16] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman. Video enhancement with the task-oriented flow, 2017. arXiv preprint arXiv:1711.09078

[17] Choi, M., Kim, H., Han, B., Xu, N., & Lee, K. M. (2020). Channel Attention Is All You Need for Video Frame Interpolation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 10663-10671. https://doi.org/10.1609/aaai.v34i07.6693

[18] S. Wen, W. Liu, Y. Yang, T. Huang and Z. Zeng, "Generating Realistic Videos From Keyframes With Concatenated GANs," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2337-2348, Aug. 2019, DOI: 10.1109/TCSVT.2018.2867934.

[19] Z. Liu, R. A. Yeh, X. Tang, Y. Liu and A. Agarwala, "Video Frame Synthesis Using Deep Voxel Flow," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4473-4481, DOI: 10.1109/ICCV.2017.478.

[20] X. Chen, W. Wang and J. Wang, "Long-term video interpolation with bidirectional predictive network," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, DOI: 10.1109/VCIP.2017.8305029.

[21] X. Chen, W. Wang and J. Wang, "Long-term video interpolation with bidirectional predictive network," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, DOI: 10.1109/VCIP.2017.8305029.

[22] Jean Bégaint, Franck Galpin, Philippe Guillotel, Christine Guillemot. Deep frame interpolation for video compression. DCC 2019 - Data Compression Conference, Mar 2019, Snowbird, United States. pp.1-10, ff10.1109/DCC.2019.00068ff. ffhal-02202172f

[23] Liu, Xiaozhang& Liu, Hui & Lin, Yuxiu. (2020). Video frame interpolation via optical flow estimation with image inpainting. International Journal of Intelligent Systems. 35. 10.1002/int.22285.

[24] Liu, Yu-Lun& Liao, Yi-Tung & Lin, Yen-Yu & Chuang, Yung-Yu. (2019). Deep Video Frame Interpolation Using Cyclic Frame Generation. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 8794-8802. 10.1609/AAAI.v33i01.33018794.

[25] T. Jayashankar, P. Moulin, T. Blu and C. Gilliam, "Lap-Based Video Frame Interpolation," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4195-4199, DOI: 10.1109/ICIP.2019.8803484.

[26] Niklaus, S., & Liu, F. (2020). Softmax Splatting for Video Frame Interpolation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5436-5445.