Hybrid Approach for Video Interpolation

R. Geetharamani
Professor, CEG, Anna University
Chennai, Tamil Nadu

Eashwar P.
B.Tech, CEG, Anna University
Chennai, Tamil Nadu

Srikarthikeyan M.K.
B.Tech, CEG, Anna University
Chennai, Tamil Nadu

Varoon S.B.
B.Tech, CEG, Anna University
Chennai, Tamil Nadu

ABSTRACT

Video interpolation is a form of video processing in which intermediate frames are generated between existing ones by means of interpolation. The research aims to synthesize several frames in the middle of two adjacent frames of the original video. Interpolation of the frames can be done in either static or dynamic mode. The dynamic approach identifies the number of frames to be interpolated between the reference frames. The reference frames are passed onto three deep learning networks namely Synthesis, Warping, and Refinement where each network performs different functionality to tackle blurring, occlusion and arbitrary non-linear motion. Synthesis is a kernel-based approach where the interpolation is modeled as local convolution over the reference frames and uses a single UNet. Warping makes use of Optical Flow information to back-warp the reference frames and generates the interpolated frame using 2 UNets. Refinement module works with the help of optical flow and warped frames to compute weighted maps which enhance interpolation using 2 UNets and GAN. The raw interpolated output of the deep learning networks yielded a PSNR of 38.87 dB and an SSIM of 97.22% was achieved. In order to further enhance the results, this research combined these deep learning approaches followed by post-processing. The raw interpolated frames are color corrected and fused to form a new frame. Color correction is the process of masking the difference between the interpolated frame and the ground truth over the interpolated frame. Fusion ensures that the maximum pixel value from each input frame is present in the fused frame. Voting is applied on the color corrected frames from the three networks and the fused frame. This voting follows a per-pixel strategy and selects the best pixel from each of the interpolated frames. Datasets used to train and test these modules are DAVIS(Densely Annotated VIdeo Segmentation), Adobe 240, and Vimeo. This hybrid interpolation technique was able to achieve the highest PSNR of 58.98 dB and SSIM of 99.96% which is better than the results of base paper that achieved a PSNR of 32.49 dB and SSIM of 92.7%.

General Terms

Video Interpolation, Deep Learning, UNet.

Keywords

Fusion, Optical Flow, Warping, Video Frame Interpolation

1. INTRODUCTION

Capturing high-resolution videos with high frame rates is a highly challenging task. They typically require professional high-speed cameras, that are inaccessible to casual users. Modern mobile device manufacturers have tried to incorporate more affordable sensors with similar functionalities into their systems, but they still suffer from the large memory requirements and high power consumption associated with these sensors.

1.1 Video Frame Interpolation

Video Frame Interpolation is a classic problem in the computer vision community with many applications such as frame rate up-conversion, slow-motion generation, and video compression. It aims to generate intermediate frames between any consecutive frames in a video sequence. In general, video frame interpolation algorithms estimate motion flow between two consecutive frames. Video Frame Interpolation (VFI) addresses this problem, by converting videos with moderate frame rates high frame rate videos in post-processing. In theory, any number of new frames can be generated between two keyframes of the input video.

1.2 Video Interpolation Methods

Frame-based interpolation approach makes use of the input from a conventional frame-based camera that records frames in a synchronous manner and at a fixed rate. There are several classes of such methods. Some of them are:

Warping based approaches combine optical flow estimation with image warping to generate the intermediate frame between two reference frame. They work under the assumptions of linear motion and brightness constancy between frames. Results are improved by making use of contextual information, visibility maps, spatial transformer networks and forward warping. Most of these approaches assume linear motion, but recent works have also been carried out assuming quadratic or cubic motions. Although these methods are capable of addressing non-linear motions, they still fail to capture arbitrary motion as they are limited by their order.

Another type of frame based interpolation is the Kernel-based approaches. It does not follow the explicit motion flow estimation and warping stages of warping-based approaches. Here, VFI is modeled as a local convolution network over the two reference key frames. By this way, this approach is more robust to light changes and motion blur.

The third approach of frame based interpolation is the phase based approaches. It depicts the VFI as a phase shift estimation problem. It consists of a neural network decoder that directly estimates the phase decomposition of the intermediate frames. However, in practice these methods can't handle large volume of motions due to the locality of the convolution kernels.

In practice, all the approaches of frame based interpolation assume linear motion models due to the absence of visual information during the blind time between frames. It may pose a fundamental limitation for purely frame-base interpolation approaches. These limitations rely on brightness and appearance constancy between frames which in turn results in limited applicability in highly dynamic scenarios such as

i. Non-linear motions between the input reference frames

ii. Changes in illumination or motion blur and

iii. New object appear or disappear in-between reference frames.

The Multi-camera approach aims to combine inputs from frame-based cameras with different spatio-temporal trade-offs like combined low-resolution video with high resolution still images. It fuses low-resolution high frame rate video with high resolution low frame rate video. This approach can find missing information from which true object motion can be estimated or reconstructed. The downside is the consumption of high power and requirement of more memory.

2. RELATED WORKS

2.1 Flow Based Method

In 2021, Minho Park et al [1], proposed a robust VFI with exceptional motion map. The proposed VFI takes into account the exceptional motion map that contains the location and intensity of the exceptional motion. The proposed method consists of three parts, which are optical flow based frame interpolation, exceptional motion detection, and frame refinement. The optical flow based frame interpolation predicts an optical flow which is used to synthesize the pre-generated intermediate frame. The exceptional motion detection detects the position and intensity of complex and large motion with the current frame and the previous frame sequence. The frame refinement focuses on the exceptional motion region of the pre-generated intermediate frame by using the exceptional motion map. This method is robust against the exceptional motions including complex and large motion and results in a Peak Signal to Noise Ratio (PSNR) of 34.71 and Structural Similarity Index Measure (SSIM) of 0.969 on UCF101 dataset.

In 2021, Jinbo Xing et al [2], proposed a generic motion model for VFI based on flow-aware synthesis. The model uses time as a control variable to interpolate multiple intermediate frame with complex non-linear motion. Adaptive Flow Prediction technique is used to approximate the complex motion in video. It predicts the optical flow information from the reference frames. Instead of directly warping the reference frames with the flow, which results in blurred frame, it uses a pyramid context extractor to extract multi-scale contextual features from the reference frame. This is warped with the reference frame using a forward warping layer. The results are sent to a frame synthesis network which produces a residual map between ground truth and average blending of the frames. The model predicts the interpolated frame by summing the residual map and the average blending of the frames. The experiments was performed on Nvidia Titan RTX Graphics Processing Unit. It gave a PSNR of 33.69 and SSIM of 0.9703 on Vimeo-90K dataset.

In 2021, Bo Yan et al [3], proposed a fine grained motion estimation approach for VFI. It consisted of 2 strategies: multi-scale coarse-to-fine optimization and multiple motion features estimation. The multi-scale coarse-to-fine optimization strategy is used to refine optical flows and weight maps which are used to synthesize the target frame. The multiple motion features estimation strategy aims to provide fine-grained motion features by generating multiple optical flows and weight maps. A CNN with three refinement scales and four motion features is used to synthesize the interpolated frames. The three refinement scales are three sub-networks that are combined and optimized in the training process. Input frames, optical flow and fusion weight produced by previous sub-network and warping result are given as input to each sub-network and its output include optical flow and fusion weights. The experiments were conducted on an Intel i7 CPU and Nvidia GTX 1080Ti Graphics Processing Unit. It resulted in a PSNR of 34.70 and SSIM of 0.9612 on Vimeo-90K dataset.

In 2021, Yong-Hoon Kwon et al [4], proposed Direct video frame interpolation with multiple latent encoders. This method is simple but effective video interpolation framework that can be applied to various types of videos including conventional videos and 360° videos. It predict the latent feature of an intermediate frame, through the latent feature encoders between encoder and decoder networks, without explicitly computing optical flow or depth maps. The latent feature encoders take latent features of input images and then predict the latent feature of a target image, i.e. an intermediate frame. Then the decoder network reconstructs the target image from the latent feature. Multiple latent encoders framework consists of CNN, and it is therefore end-to-end trainable from scratch without requiring additional information except for consecutive frames. This proposed method performs interpolation in latent domain and can be applied on various types of input data. Results for Vimeo-90K dataset show PSNR of 34.96 and SSIM of 0.9753 respectively.

In 2020, Minho Park et al [5], proposed VFI with Exceptional Motion aware synthesis. The method used two deep learning modules - exceptional motion detection and frame interpolation with refined flow. The motion detection module detects the position and intensity of exceptional motion patterns in current frame given the past frame sequence. The flow refinement module refines the pre-estimated bidirectional optical flow using exceptional motion information. Thus, refined optical flows are obtained and intermediate frame is produced by warping. The proposed method improves the quality of the synthesized intermediate frame by making the optical flow robust against exceptional case of motion. The model was trained on a system with Intel core i7, 32 GB of memory and Nvidia GeForce GTX 1080 Ti. It resulted in a PSNR of 35.90 and SSIM of 0.970 on UCF101 dataset.

2.2 Warping Based Approach

In 2020, Avinash Paliwal et al [6], proposed VFI using Deep Slow Motion Video Reconstruction with Hybrid Imaging system. The main theme is the Two-Stage Deep Learning system consisting of Alignment and Appearance estimation that reconstructs high resolution slow motion video from the hybrid video input (main video with low frame rate and high spatial resolution along with an auxiliary video with high frame rate and low spatial resolution). It has a flow estimation system that utilizes two videos to generate high resolution flows in large motion. A context and occlusion aware appearance estimation network which blends the two warped key frames and minimizes warping artifacts is introduced. The results are demonstrated using two real dual camera rigs with small baseline. It address the lack of temporal information in the low frame rate input video by coupling it with a high frame rate video with low spatial resolution. It resulted in a Learned Perceptual Image Patch Similarity (LPIPS) of 0.1332 and SSIM of 0.865 on Middlebury and Adobe-240 dataset.

In 2020, Joi Shimizu et al [7], proposed High Efficiency Video Coding (HEVC) with deep learning based frame interpolation. This model is based on novel video compression method which incorporates deep learning based frame interpolation into HEVC which is the current video compression standard. An input video is compressed with HEVC. Next, a new video is created from the odd frames of the input video. This new video is compressed with HEVC again. After the compression, even frames of the original video are interpolated using a deep learning based frame interpolation algorithm. The model further utilizes the residual image to improve the quality of interpolated frames. Finally, the compressed residual images are added to predicted frames and a new video is created. Depth Aware video frame INterpolation (DAIN) performs well for the prediction of even frames and this method outperforms HEVC in some sequences.

By making use of residual frames, the model was found to be more stable.

2.3 Synthesis Based Techniques

In 2021 Xiangling Ding et al [8], proposed Detection of deep VFI via learning dual-stream fusion CNN in the compression domain. Here a hybrid neural network has been implemented to localize the deep interpolated frames by learning spatio-temporal representations from the residual and motion vector information in the compression domain. First the residual and motion vector of motion regions are maintained by an intra-prediction constraints. Then, inherent tampering traces are further highlighted through subtracting the estimate of the residual or motion vector by virtue of residual modulation or Motion Vector (MV) refinement network. And at the end an attention-based dual-stream network is designed to jointly learn discriminative representations from the enhancement traces. Results for Vimeo-90K dataset show F1 score of 80.82 respectively.

In 2020, Jiankai Zhuang et al [9], proposed VFI with a Lightweight Network model using spatial pyramids. Here, a cascaded network called Spatial Pyramid Frame Interpolation Network (SPFIN) is used to break down frame interpolation into several small and easy-to-learn problems. Part of the network only estimates a minor update based on results from previous level and it can be designed as compact as possible. At each level, a two-step procedure consisting of approximation of bi-directional flow and refinement is conducted. Moreover, SPFIN could generate frame at any time by introducing time variable into modeling. This approach has the smallest model size (9.0 MB) with better and comparable performance to existing state of the art models. The evaluation metrics used here are PSNR, SSIM, and Interpolation Error (IE). It resulted in a PSNR of 33.07, SSIM of 0.937 and IE of 7.88 on UCF-101 dataset.

In 2018, Chenguang Li et al [10], proposed VFI based on Multi-Scale CNN and Adversarial Training. It synthesizes the interpolated frames with favorable quality and visual experience. It uses a combination of loss function, including a Wasserstein generative adversarial network loss with gradient penalty. It has a slim generator network structure in order to meet the real-time interpolation requirement as much as possible that paves way for less parameters, which could be beneficial to video processing tasks in future works. This model is capable of handling different range of movement to obtain sharp results, and the adversarial training keeps the generated frames natural. It contains a perceptual loss, which can improve the visual quality of the interpolated frames. It takes a level-by-level training strategy which can speed up the convergence of network learning. It resulted in a PSNR of 34.11 and SSIM of 0.943 on UCF-101 dataset.

In 2017, Xiongtao Chen et al [12], proposed VFI using Long-Term Video Interpolation with Bidirectional Predictive Network (BiPN). It attempts to speculate or imagine the procedure of an episode and further generate multiple frames between two non-consecutive frames in videos. It presents a novel deep architecture called BiPN that predicts intermediate frames from two opposite directions. The bidirectional architecture allows the model to learn scene transformation with time as well as generate longer video sequences. A joint loss is composed of clues in image and feature spaces and adversarial loss is designed to train the model. The network consists of a bidirectional encoder-decoder that predicts the future forward from start frame and predicts the past backward from end frame at the same time. It resulted in a PSNR of 31.4 and SSIM of 0.94 on UCF-101 dataset.

3. APPROACH

The system architecture of the paper is shown in Figure 1.

3.1 Dynamic Estimation of Frames

Based on input reference framesestimation of the number of frames to be interpolated is carried out. PSNR is calculated between these input frames and based on the PSNR value along with various different frames analysis number of frames to be interpolated is determined. In a given sequence some frames have a huge differences between them so static interpolation won’t work well. To overcome the inequality between these sequence of frames dynamic approach is used. The number of frames to be interpolated will be the output.

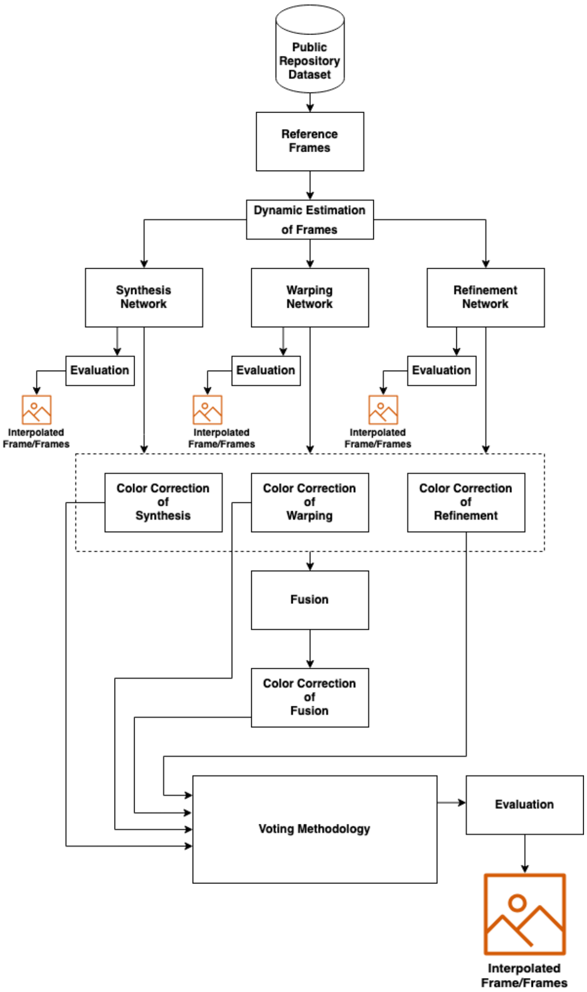
Public Repository Dataset
Reference Frames
Dynamic Estimation of Frames
Synthesis Network
Warping Network
Refinement Network
Evaluation
Evaluation
Evaluation
Interpolated Frame/Frames
Interpolated Frame/Frames
Interpolated Frame/Frames
Color Correction of Synthesis
Color Correction of Warping
Color Correction of Refinement
Fusion
Color Correction of Fusion
Voting Methodology
Evaluation
Interpolated Frame/Frames

Fig 1: System Architecture

3.2 Synthesis Network

The synthesis network, Figure 2, is adapted from the kernel based approach of interpolation. The sliding window of kernel, also called as filter, is applied on images with dot-product gives a feature map. It can be interpreted as something that gets activated at certain visual features. The deeper the level, the more elusive the representation of feature maps will be. ReLU(Rectified Linear Unit) is used as activation function, which helps prevent from having vanishing gradients through the network.

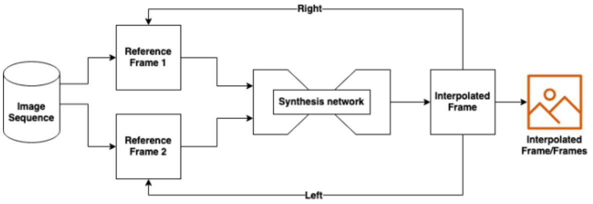

Fig 2: Synthesis Network

The whole network(U-Net) is an encoding and decoding structure which is five levels deep and each layer is connected with its parallel layer as shown in. Adam (Adaptive moment estimation) optimizer is used, which is a manifestation of stochastic gradient-based optimization. The input reference frames are concatenated and passed into the synthesis network to obtain the interpolated frames dynamically. The interpolation is done based on the threshold PSNR between the reference frames in the input sequences.

3.3 Warping Network

The Warping Network as shown in Figure 3 predicts the intermediate frame using the 2 reference frames I_0 and I_1, I_t. This can be done when the optical flow fields from I_t to I_0 ($F_{t\to 0}$) and I_t to I_1 ($F_{t\to 1}$) are known. The backward warping function makes use of bi-linear interpolation. The two parameters that control the contribution of the reference frames are: Temporal consistency and Occlusion. In temporal consistency, if a frame has to be interpolated close to T = 0, then I_0 contributes more to I_t. Similar property holds for I_1 as well. Another property is that if a pixel is visible at T=t, it is likely to be visible in at least one of the reference frames. Thus occlusion problem can be handled by using Visibility maps $V_{t\leftarrow 0}$ and $V_{t\leftarrow 1}$. A value of 0 means the pixel is occluded and 1 means the pixel is visible. However computing the flow fields $F_{t\to 0}$ and $F_{t\to 1}$ is a difficult task. To address this problem, the bidirectional flow $F_{0\to 1}$ and $F_{1\to 0}$ can be computed using the UNet which takes the reference frames as inputs. The intermediate frames are back warped using the bi-directional flow and reference frames.The timestamp values are calculated based on the number of frames to be calculated. Using the timestamp values and the bidirectional flow, the flow fields are synthesized.This works well for smooth regions but not for motion boundaries. To solve this issue, another UNet is used.

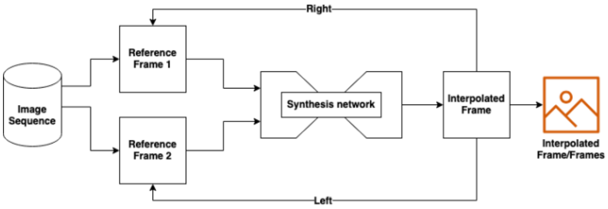

Fig 3: Warping Network

This takes the reference frames, bidirectional flow, flow information and intermediate warped frames as input and produces the refined flow and visibility maps as output. Using this refined flow and the reference frames, the required frames can be interpolated.

3.4 Refinement Network

With given reference frames, first estimation of optical flow from 0 to t and t to 1 where t ranges from 0 to 1 by passing these frames in UNet is done. Optical flow works on individual pixel level and tries to estimate how each pixel's color changes on the image plane over time. After estimating the flow, warping is done by combining optical-flow results and reference frames. With optical flow estimation and warping the outputs are passed into UNet to generate weight maps. It is used to capture information about lighting changes and occlusion.

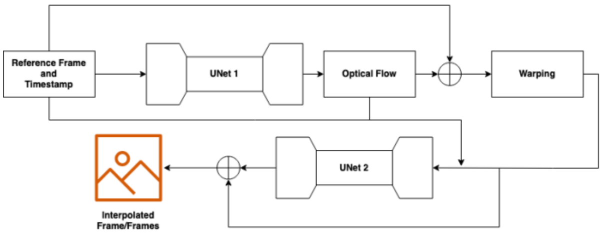

Fig 4: Refinement Network

Weight maps are combined with warped frames to get interpolated frame as shown in Figure 4. Finally interpolated frame is passed into GAN(Generative Adversarial Network) network which is shown in for further refinement. A typical loss functions like MSE(Mean Square Error) can only provide feedback about the pixel-wise differences between the generated samples and the real samples but it does not tell the generator, in this caseit's UNet, whether the generated samples are realistic or not. GAN solves this by introducing a discriminator network whose output serves as a metric to determine how realistic the prediction is.

3.5 Color Correction

Frames generated from Networks of Synthesis, Warping, Refinement and their Fusion are the inputs in this module. Color correction is the initial step of the post-processing. Output Images from the networks of Synthesis, Warping, and Refinement are the individual inputs for the color correction module. The difference between the ground truth and the input frame is estimated. The pixel wise differed values from the ground truth image and the input frames are analyzed to set the limit for color correction. The difference array is used as the mask over the input frame to synthesize the color corrected frame. The color corrected frame yields better results compared to the raw outcomes of the various networks used for interpolation.

3.6 Fusion

The interpolated frames of Synthesis, Warping and Refinement modules are given as input. The color corrected frames of the synthesis, warping and refinement are passed onto the fusion module as shown in Figure 5. The objective of this module is to combine the three interpolated frames into a single frame. This ensures that the best pixels from each input frame is present in the fused frame. This fused frame is further color corrected with respect to its ground truth in the Color Correction module to further improve its quality.

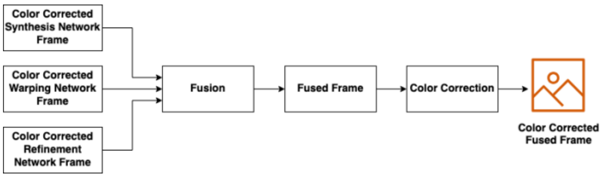
Color Corrected Synthesis Network Frame
Color Corrected Warping Network Frame
Color Corrected Refinement Network Frame
Fusion
Fused Frame
Color Correction
Color Corrected Fused Frame

Fig 5: Fusion

3.7 Voting Methodology

The Voting module as shown in Figure 6 performs a pixel wise comparison of the aforementioned four frames. If the pixel values at a particular location of all the input images are same, then no change is made to that pixel. The same is done if majority of the pixel values are same in the input image. If the pixel values differ, then the candidate for the pixel value is determined using multiple approaches to find the best result.

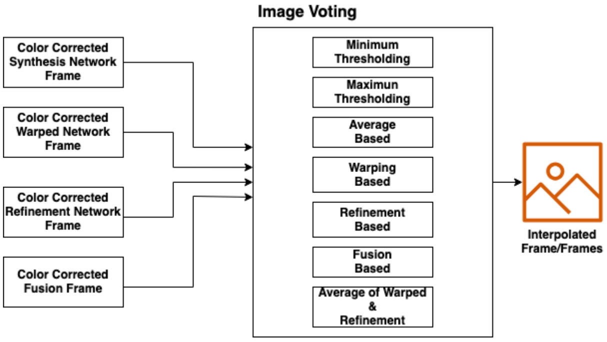
Image Voting
Color Corrected Synthesis Network Frame
Color Corrected Warped Network Frame
Color Corrected Refinement Network Frame
Color Corrected Fusion Frame
Minimum Thresholding
Maximun Thresholding
Average Based
Warping Based
Refinement Based
Fusion Based
Average of Warped & Refinement
Interpolated Frame/Frames

Fig 6: Voting Methodology

These approaches include:

a. Minimum thresholding - Finding the minimum pixel value among the four existing values.

b. Maximum thresholding - Finding the maximum pixel value among the four existing values.

c. Average Based - Finding the average of the four existing pixel values.

d. Warping Based - The corresponding pixel value in the warped frame is substituted here.

e. Refinement Based - The corresponding pixel value in the refinement frame is substituted here.

f. Fusion based - The corresponding pixel value in the fused frame is substituted here.

g. Average of Warping and Refinement - Finding the average pixel value of the warped and refinement frames.

These resultant frames of various approaches are compared and the best resultant frame is given as output.

3.8 Evaluation of Interpolated Frames

PSNR is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Ground Truth (GT) and interpolated frame are used to calculate the Mean Square Error (MSE). PSNR value is then calculated using the MSE. SSIM is the metric that quantifies image quality degradation caused by Image processing. The prerequisite for calculating SSIM is that both the ground truth and interpolated images are converted into gray-scale images.

4. RESULTS AND DISCUSSION

Various experiments were performed as a part of this research which involved interpolating one or more intermediate frames and evaluating them against the ground truth. Theconsolidated results of the experiments using different scenarios are shown in Table.1 and Table 2.

Table.1 Performance Analysis

Experiments	Synthesis				Warping				Refinement				Fusion			
	Non-color corrected		Color corrected		Non-color corrected		Color corrected		Non-color corrected		Color corrected		Non-color corrected		Color corrected	
	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)
Balloon	33.5	90.82	50.8	99.16	38.87	97.22	54.92	99.78	41.09	98.68	**55.7**	**99.80**	34.9	92.90	54.49	99.68
Bear	27.9	82.97	47.3	97.20	33.33	90.94	47.69	99	33.6	91.86	48.02	**99.23**	31.12	76	**50.36**	98.80
Boat	29.8	89.74	46.2	97.99	33.86	92.08	**50.83**	99.31	33.6	92.50	47.5	**99.45**	31.57	85.10	49.13	98.38
Swing	29.9	88.78	44.3	95.93	33.72	90.37	42.98	94.68	34.3	99.25	**48.5**	**99.30**	30.93	87.98	48.23	98.24

Table.2 Voting Results

Experiments	Minimum		Maximum		Average		Fusion		Refined Warped		Warping		Refinement	
Metrics	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)	PSNR (dB)	SSIM (%)

Bear	39.87	97.25	40.4	97.80	32.57	24.53	37.84	95.34	45.67	63.49	**47.39**	99.03	44.77	**99.11**
Swing	45.65	97.03	48.31	98.51	44.02	94.40	44.63	96.20	49.27	96.15	49.69	99	**50.54**	**99.52**
Boat	49.5	98.67	48.83	98.35	46.13	96.64	46.56	97.04	54.21	97.75	**55.55**	99.77	54.99	**99.83**
Balloon	52.4	99.49	53.65	99.66	50.4	98.79	51.04	99.25	56.5	99.17	**58.98**	**99.96**	56.92	99.86

4.1 Fusion

This experiment involves the raw interpolated frames from the networks of Synthesis, Warping and Refinement. These three frames are combined to produce the Fused frame. All these four frames are shown in Figure 8. The Synthesis frame has obtained a PSNR of 33.50 dB and SSIM of 90.82%. The Warping frame has achieved a PSNR of 38.87 dB and SSIM of 97.22%, while the Refinement frame has got a PSNR of41.09 dB and SSIM of 98.68%. Their fused frame has achieved a PSNR of 34.90 dB and SSIM of 92.20%. It can be observed that the Refinement frame performs better when the raw interpolation results are compared.These results are improved in the following experiments.

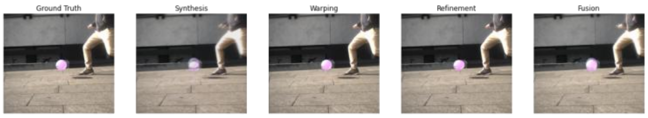

Fig 8: Interpolated Frames of Synthesis, Warping, Refinement and Fusion Modules

4.2 Color Correction

This approach is based on the difference in pixel values between the ground truth and the interpolated frames. Mask is created using the difference in pixels and applied over the interpolated frames. This technique greatly enhances the image quality and achieves better evaluation metrics. Color correction is used at different levels in this project. Raw interpolated images from the networks of Synthesis, Warping, Refinement and their Fused output are color corrected. Those images are shown in Figure 9. The increase in PSNR ranges from 10 dB to 20 dB when compared to the non-color corrected frames. The fused color corrected Balloon image obtained a PSNR of 54.49 dB and SSIM of 99.68% when compared to the non-color corrected version which had a PSNR of 34.90 dB and SSIM of 92.90%.

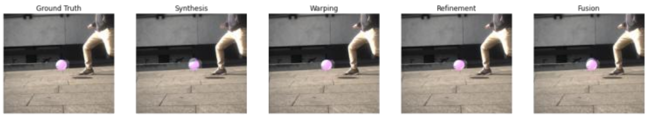

Fig 9: Color Corrected Frames

4.3 Voting Methodology

The Synthesis, Warping, Refinement and Fusion module's color corrected images were taken for this experiment. Several voting approaches were carried out. These approaches were based on minimum thresholding, maximum thresholding, fusion based, Refinement based, warping based and average based and Refined-Warp based (the average of Warping and Refinement frame) as shown in Figure 10. Out of these multiple approaches, the Warping based one obtained the highest PSNR value of 58.98 dB and SSIM values of 99.96%.

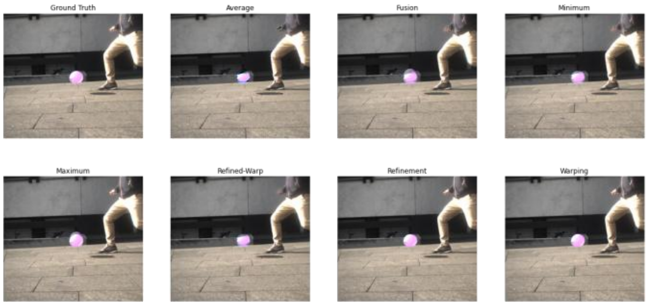

Fig 10: Different approaches of Voting Methodology

4.4 Application in Various Scenario

The balloon falling experiment involved the linear fall of the water balloon and the man's leg movement. In order to evaluate the efficiency of this hybrid approach, testing with other datasets were also carried out.

These experiments involved interpolating intermediate frames for different scenarios like a bear moving in the forest, a boat moving forward in the sea, a little girl swinging in the grassland and a real-time video of door closing. In majority of the cases, it is observed that the frames generated by the voting methodology yielded better results, in-particular the warping, refinement and refined-warp based voted frames.

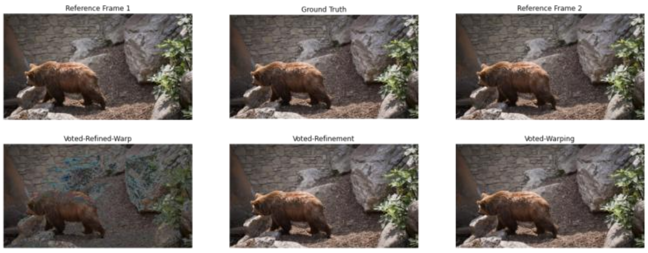

Fig 11: Ground Truth, Reference frames and Voting results of Bear from Vimeo Triplet dataset

As shown in Figure 11, for the bear movement, the entire frame is in motion since the camera angle is modified. Here the refined-warp method produced the lowest metrics while the highest metrics were obtained in warping method. The highest PSNR of 50.36 dB and SSIM of 99.23% was obtained.

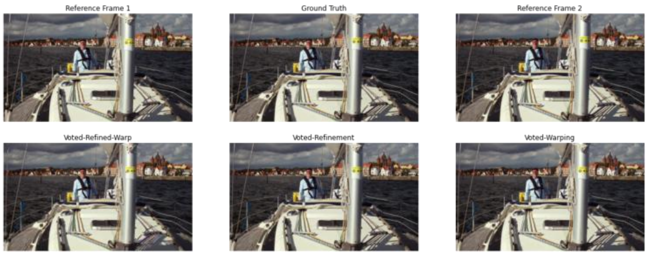

Fig 12: Ground Truth, Reference frames and Voting results of Boat from Vimeo Triplet dataset

As shown in Figure 12, for the boat movement, the boat is in motion towards the camera. In refinement method, there were minimal distortions in the ropes of the boat which were not present in warping method. The highest PSNR of 55.55 dB and SSIM of 99.83% was achieved.

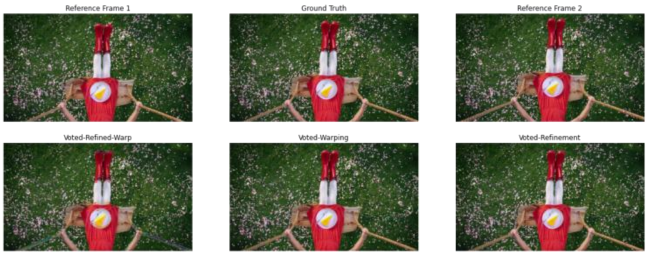

Fig 13: Ground Truth, Reference frames and Voting results of Swing from Vimeo Triplet dataset

As shown in Figure 13, for the swing movement, the background contained high detailing which was not accurately interpolated by warping method but the details persisted in the refinement method. The highest PSNR of 58.98 dB and SSIM of 99.96% was obtained.

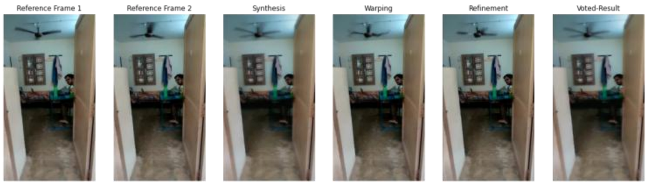

Fig 14: Reference frames and Voting results of Realtime Scenario

As shown in Figure 14, for the real-time scenario, there were two motions such as rotation of fan blades and closing of the door. The fan rotation is poorly interpolated in both warping and refinement methods but in synthesis method it was interpolated comparatively well. The door movement is interpolated accurately in warping and refinement methods. The visually better image is obtained using the voting methodology after the fusion of the raw interpolated frames.

5. CONCLUSION

VFI is carried out by synthesizing intermediate frames between the reference frames using the Warping, Synthesis and Refinement Network followed by fusion module. Color correction is done for all three network outputs and passed into the fusion module which is also color corrected. The post-processing involves various voting methodology approaches. By doing so, this interpolation technique was able to achieve the highest PSNR of 58.98 dB and SSIM of 99.96% which are the evaluation metrics. The model's working was understood by performing several experiments with publicly available datasets along with real-time data.

As a part of learning through experimenting, some findings were made and flaws were identified which can be rectified in the future to increase results further high. Making use of image segmentation deep learning models like Edge Detection Segmentation, Image Segmentation based on Clustering, Mask R-CNN with UNet can be done. This research requires a huge training duration which is one of the flaw but higher the complex model better results can be obtained.

6. REFERENCES

[1] Minho Park, Hak Gu Kim, Sangmin Lee and Yong Man Ro. Robust Video Frame Interpolation with Exceptional Motion Map. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 31(2):754–764, 2021.

[2] Jinbo Xing, Wenbo Hu, Yuechen Zhang and Tien-Tsin Wong. Flow-aware synthesis: A generic motion model for video frame interpolation. Computational Visual Media, 7(3):393–405, 2021.

[3] Bo Yan, Weimin Tan, Chuming Lin and Liquan Shen. Fine-Grained Motion Estimation for Video Frame Interpolation. IEEE TRANSACTIONS ON BROADCASTING, 67(1):174–184, 2021.

[4] Yong-Hoon Kwon, Ju Hong Yoon and Min-Gyu Park. Direct Video Frame Interpolation with Multiple Latent Encoders. IEEE Access, 9:32457–32466, 2021.

[5] Minho Park, Sangmin Lee and Yong Man Ro. Video Frame Interpolation via Exceptional Motion-Aware Synthesis. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1958–1962, 2020.

[6] Avinash Paliwal and Nima Khademi Kalantari. Deep Slow Motion Video Reconstruction with Hybrid Imaging System. IEEE Transactions Pattern Analysis And Machine Intelligence, 42(7):1557–1569, 2020.

[7] Joi Shimizu, Zhengxue Cheng, Heming Sun, Masaru Takeuchi and Jiro Katto. HEVC Video Coding with Deep Learning Based Frame Interpolation. IEEE 9th Global Conference on Consumer Electronics (GCCE), pages 433–434, 2020.

[8] Xiangling Ding, Yifeng Pan, Qing Gu, Jiyou Chen, Gaobo Yang and Yimao Xiong. Detection of Deep Video Frame Interpolation via Learning Dual-Stream Fusion CNN in the Compression Domain. IEEE International Conference on Multimedia and Expo (ICME), pages 1–6, 2021.

[9] Jiankai Zhuang, Zengchang Qin, Jialu Chen and Tao Wan. A Lightweight Network Model for Video Frame Interpolation using Spatial Pyramids. International Conference on Image Processing (ICIP), pages 543–547, 2020.

[10] Chenguang Li, Donghao Gu, Xueyan Ma, Kai Yang, Shaohui Liu and Feng Jiang. Video Frame Interpolation based on Multi-Scale Convolutional Network and Adversarial Training. IEEE 3rd International Conference on Data Science in Cyberspace, pages 553–560, 2018.

[11] Xiongtao Chen, Wenmin Wang and Jinzhuo Wang. Long-Term Video Interpolation with Bidirectional Predictive Network. VCIP,St Petersburg, U.S.A, pages 1–4, 2017.