Slow-Motion Video Reconstruction

THAMMINENI DAYAKAR1, MENTA VIJAYABHASKAR2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ Slow Motion is an application of Computational Photography to boost the slow-motion video capture capabilities of cameras by exploiting dual camera setups. shooting high-resolution videos at extremely high frame rates requires sophisticated high-speed cameras. Slow motion cameras require pricey cameras. However, due to the improvements and deep research in computer vision domain, we now have several straightforward and inexpensive approaches for getting more frames per second (FPS), employing Deep Learning and Video Frame Interpolation. Video frame interpolation automatically recovers missing video frames between the original videos which delivers a better watching experience and it also raises the frames per second. The two-stage deep learning system comprises of alignment and appearance estimation using CNN that reconstructs high-resolution slow-motion video from the hybrid video input. The model trained on synthetically generated hybrid videos and exhibited high-quality results on a number of test settings. The project focuses on the main and auxiliary films as input and creates a movie with high spatial resolution and high frame rate videos as output

1.INTRODUCTION

It's tough to catch everything that happens in a split second, whether it's a pop of champagne or a bolt of lightning. You can play back the resulting video in slow motion because it was shot with a high frame rate camera. Slow motion videos are getting more and more popular since smartphones have the ability to record videos with high frame rates. In order to achieve a higher frame rate, the spatial resolution of these cameras must be sacrificed. Even while Nokia 6.1 is capable of recording high-definition films at 30 frames per second, it is not capable of recording 480p videos at 120 frames per second, for example. If you're looking for the most advanced high-speed cameras, you'll have to fork out a lot of money. In addition, many of the times we would like to slow down are unpredictable, thus they are captured at conventional frame rates. For mobile devices, it's impracticable to record everything at fast frame rates because enormous memory capacity and significant battery consumption are both required.

That's why slow-motion footage from existing videos is of huge interest. Video interpolation can be used to create smooth view transitions as well as to increase

the frame rate of conventional videos. Self-supervised learning also offers exciting new uses, using it to learn optical flow from unlabeled videos [1].

Tasks in Slow-Motion Conversion:

There are two main tasks in converting a normal video to slow-motion video. They are listed below:

1. Video Frame Interpolation
2. Video Reconstruction

Video Frame Interpolation, also known as in-betweening, is the process of generating intermediate frames between two consecutive frames in a video sequence. This is an important technique in computer animation, where artists draw keyframes and let software interpolate between them. With the advent of high frame rate displays that need to display videos recorded at lower frame rates, in-betweening has become important in order to perform frame rate up-conversion. Computer animation research indicates that good in-betweening cannot be obtained based on linear motion, as objects often deform and follow nonlinear paths between frames. The goal of Video Frame Interpolation is to synthesize several frames in the middle of two adjacent frames of the original video. Video Frame Interpolation can be applied to generate slow motion video, increase video frame rate, and frame recovery in video streaming.

Video Reconstruction is creating/adjusting the video from the frames that are randomized or jumbled. For alignment, we propose to compute flows between the missing frame and two existing frames of the main video by utilizing the content of the auxiliary video frames. For appearance estimation, we propose to combine the warped and auxiliary frames using a context and occlusion aware network. We train our model on synthetically generated hybrid videos and show high-quality results on a variety of test scenes. Construct the video from the frames produced by the interpolation and find the exact flow of the video through this reconstruction process.

We propose a system that takes video as input that applies video frame interpolation and reconstructs the video from the created intermediate frames. We developed a high-quality variable-length multi-frame interpolation method that can interpolate a frame at any arbitrary time step between two frames. Our main idea is to warp the input two images to the specific time step and then adaptively fuse the two warped images to generate the intermediate image, where the motion interpretation and occlusion reasoning are modeled in a single end-to-end trainable network. we develop a CNN architecture that directly estimates asymmetric optical flows and weights from an unknown intermediate frame to two input frames. We use this to interpolate the frame in-between. Existing techniques either assume that this flow is symmetric or use a symmetric

approximation followed by a refinement step. For nonlinear motion, this assumption does not hold, and we document the effect of relaxing it. We rely on the fact that interpolated frames can be used to estimate the original frames by applying the method again with the in-between frames as input. The similarity of reconstructed and original frames can be considered a proxy for the quality of the interpolated frames. For each frame we predict, the model is fine-tuned in this manner using the surrounding frames in the video

2.LITERATURE SURVEY

Numerous fields of study have seen significant progress thanks to machine learning methods based on artificial neural networks (ANNs). These advancements have been expedited by the adoption of Deep Learning (DL) methodologies, which are the multi-layered structure of ANNs, and the advancements in GPU technology. The state-of-the-art methodologies in several disciplines have been greatly surpassed by DL approaches, including, but not limited to: object identification; image processing; computer vision; speech recognition; natural language processing (NLP); character recognition; and signature validation. Although McCulloch and Pitts' 1943 ANN served as the basis for DL, its genuine popularity really grew in 2012. Due to the fact that they continuously obtain a local optimal solution, multilayer neural networks have proved ineffective. For the past few years, interest in multi-layered neural networks has waned because the processing capacity of large datasets has grown so rapidly. When it comes to training deep learning, Hinton and his colleagues in 2006 advocated two stages: pretraining and fine-tuning. As a result of this, interest in Digital Literacy began to rise. When Krizhevsky et al. improved the ImageNet competition's Top-5 error rate from 26.2% to 15.3% in 2012, it was a major breakthrough in object recognition. DL's popularity has risen as a result of this accomplishment in the academic community. In addition to academics, several technological businesses assist the development of DL techniques.. Researchers working in the field of deep learning (DL) are able to use frameworks established by businesses like Google, Facebook, Microsoft, and NVIDIA as open source software. Many layers, including Fully Connected (FC), Dropout and Pooling, are responsible for the success of Deep Learning. DL uses the backpropagation algorithm to show how a machine's internal parameters should be changed to compute the representation between layers in order to identify intricate structure in massive data sets. It is possible to find DL architectures in the literature, including Deep Neural Network (DNN), Convolutional Neural Network (CNN), Deep Belief Network (DBN), Sparse Auto-Encoder (SAE), and Recurrent Neural Networks (RNN). DL models differ

from one another, despite the fact that their basic architectures are the same [11].

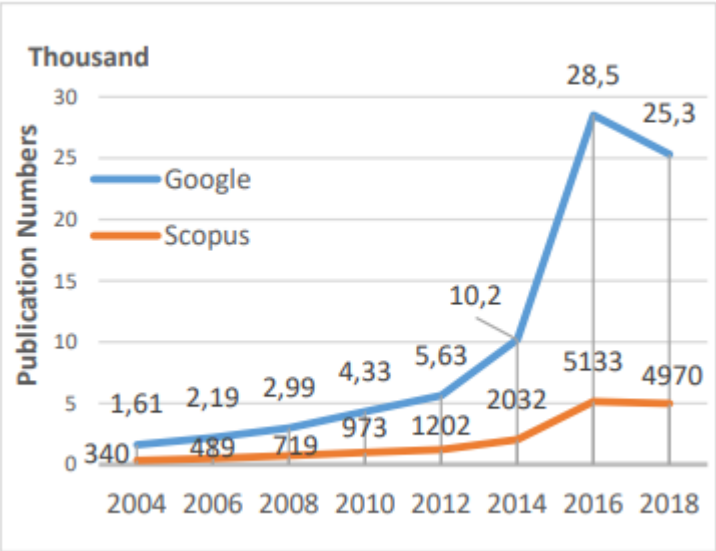


Fig. 1. Number of published deep learning articles by year. The numbers of articles were obtained from the search results on Scopus and Google Scholar with the query of 'Deep learning' [11].

On the same level as an ANN, DNNs have three main components: the input, the hidden layers, and the output. DNNs have more hidden layers than ANNs, which influences the algorithm's depth in DNNs. ANNs have only one hidden layer. Data is fed into the input layer and the output layers are used to calculate all values progressively. Data from the previous layer is fed into the following one. The sum of the multiplication of the input and the weight for each unit in the current layer yields the output values. Finally, the output values of the layer are computed by applying a nonlinear function such as a rectified linear unit (ReLU), hyperbolic tangent, or sigmoid. The final layer of output data contains slightly more abstract representations of the raw input data than the first layer. As a result, the data may be classified more accurately. More researchers in various domains, including as autonomous cars, medical image processing, big data, natural language processing (NLP), and signature recognition, are likely to turn to DL in light of the rapid growth in its application shown in Figure 1 [11]. The benchmark datasets UCF101, Middlebury, and Vimeo-90k, as well as some other relevant datasets as Adobe240 KITTI or DAVIS, are widely discussed in the interpolation domain and will be taken into consideration while performing video frame interpolation. Three consecutive frames from a video serve as one input unit in the datasets' triplets, which are commonly found in the datasets. Depending on their learning architecture, different writers use these datasets in different ways for training and evaluation [12].

User-uploaded YouTube videos form the basis of UCF101, which is a dataset including real action sequences. It's divided into 101 different sections, each with its own set of subcategories. There are 101 categories of actions in the

UCF101 dataset, compared to the preceding UCF50 dataset's 50. The UCF101 dataset has a total of 13320 movies in 101 different categories, making it the most diversified collection in terms of actions.

When evaluating video frame interpolation techniques, the Middlebury dataset is an optical flow benchmark dataset. Middlebury datasets are divided into two groups. To begin, there is the Other set, which contains the middle frames containing the ground truth, and then there is the Evaluation set, which contains the frames containing the ground truth but does not publish the results to the benchmark website for evaluation. In this collection, the image resolution is 640 x 480 pixels.

Video clips of 720p or higher are included in the Vimeo-90k dataset, which includes more than 89,000 clips from the Vimeo video-sharing platform. More than 51,000 triplets for training are available for use in the Vimeo-90k dataset. 3 successive video frames with a resolution of 448256 pixels make up each triplet. [12].

3.PROPOSED SYSTEM

As a first step, we present a CNN architecture that can directly estimate asymmetric optical fluxes and weights from an unknown intermediate frame to two input frames. This is what we use to fill in the blanks between the two frames. To get an estimate of the original frames, we utilise interpolated frames as input and apply the algorithm to the in-between frames. As a follow-up, we've devised a new method for customising video networks for each frame that appears. The quality of the interpolated frames can be gauged by comparing the similarity of the reconstructed and original frames. As we predict each frame, the model is fine-tuned using the video's surrounding frames.

Frame Interpolation:

Motion estimation and frame synthesis are often the first two processes in video frame interpolation. There are many optical flow techniques that use interpolation errors as a measure for motion estimate. Frame synthesis can then be accomplished, for example, by the use of basic hole filling and bilinear interpolation and occlusion reasoning. There are a number of other ways for synthesising new frames that leverage phase decompositions of the input frames, or local per pixel convolution kernels on the input frames, to both depict motion and synthesis new images. Solve a costly optimization problem to find out where each pixel in an intermediate frame comes from in the surrounding input frames.

When it comes to training a CNN to anticipate symmetrical optical flow, the emergence of CNNs has spawned a number of innovative learning-based approaches. By interpolating the values in the input frames, they create the target frame. While other methods rely on intermediary steps, ours uses a bidirectional

flow prediction step to begin with and then extracts context maps from the input frames to forecast the end flow to the output frames.

Every second, a large amount of frames must be extracted from the video to create slow motion. It becomes rough and unwatchable if we don't record enough frames, except when we use advanced AI methodologies to envision the additional frames by using deep learning algorithms to transform 30 frames per second video into appealing 240 frames per second slow motion. This is the only way to accomplish this. The AI framework generates intermediate motion by tracing the progression of objects from one frame to the next, starting with two distinct frames. It's not quite the same as imagining a scene in your head, but it's near enough to get the job done. The process needs to be refined before it can be used to add slow-motion effects to smartphone recordings of everyday life events [12] However, when improved, the method could be used to enhance these recordings..

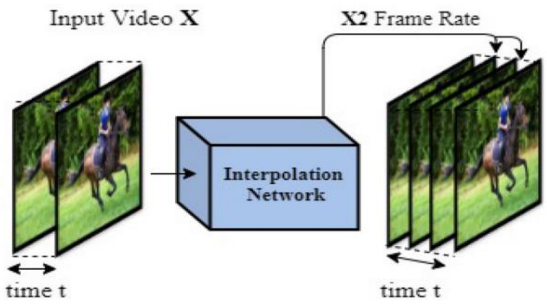


Fig. 2. General process for video frame interpolation

Video Reconstruction:

Reconstructing video from a jumbled collection of frames is an exhausting endeavour. We, on the other hand, make use of recursive frame generation to ensure that each frame is created in the correct sequence. Those frames are saved in a directory that can be immediately used in video reconstruction. Video can be edited and created with FFmpeg, a robust multimedia processing application that can handle a variety of formats. An image sequence can be used to make an animated video. When using FFmpeg, it might be tough to use because it requires a lot of understanding of the tool and its choices in order to get the results you desire. Fine-tuning the video frame interpolation can help improve the video's flow.

FFmpeg is all that is needed to assemble the pictures generated by recursive interpolation into a single video.

4.RESULTS AND DISCUSSION

Our technique is tested on a variety of datasets, including UCF101. There are numerous action sequences in UCF101, and we evaluate our method on the same frame but do not employ motion masks since we want to perform equally well over the full frame. SlowFlow and See You Again were also used to test our method.

When compared to what is currently available as far as technology goes: SlowFlow and See You Again both show that our best technique, with or without CFT, outperforms the rest. All but CyclicGen perform better than our best technique, which has the same SSIM but a lower PSNR on UCF101. The lack of access to 2 frames in all sequences by our CFT may be a contributing factor, as we hypothesise. We had to choose 1, +3 as the intermediate frame for some of the sequences because it was near the beginning of the series. In terms of visual appeal, our approach comes out on top, as illustrated in Figure 5. This could be an indication of overfitting because CyclicGen was trained on UCF101 and performed so poorly on the other datasets..

	SlowFlow		See You Again		UCF101	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DVF	-	-	-	-	34.12	0.941
SuperSloMo	-	-	-	-	34.75	0.947
SepConv $\mathcal{L}_1$	34.03	0.899	42.49	0.983	34.78	0.947
CyclicGen	31.33	0.839	41.28	0.975	**35.11**	**0.949**
Our baseline	34.28	0.903	42.50	0.984	34.39	0.946
+ asymmetric flow	34.33	0.904	42.54	0.985	34.40	0.946
+ feature loss	34.29	0.901	42.62	0.984	34.60	0.948
+ cyclic loss	34.33	0.900	42.73	0.984	34.62	0.947
+ motion loss	34.31	0.900	42.74	0.984	34.61	0.948
+ larger patches	34.60	0.907	43.14	0.986	34.69	0.948
+ CFT	**34.91**	**0.912**	**43.21**	**0.986**	34.94	**0.949**

Fig. 20. Comparison with State-of-art methods

It takes around 6.5 seconds longer for our technique to run per frame pair when we include CFT. Because we just fine-tune on 256 256 patches for 50 iterations every frame pair, this isn't affected by the size of the image. It takes 0.08 seconds to interpolate a 1920x1080 image on an NVIDIA GTX 1080 TI using our approach without CFT. When calculating many in-between frames, CFT is only required once per frame pair in the source video, therefore there is no additional overhead. To get the best results, we can only improve a proxy for interpolation quality by training for more than 50 times. The ideal number of iterations is yet to be discovered, but the quality of the pre-training, the training parameters, and the specific video all play a role. Only utilise CFT at this time if your primary concern is with interpolation quality. CFT speed improvement is an issue that deserves additional study. As an alternative, training a network to anticipate weight changes or learning the outcomes of CFT may be an option for getting similar results..

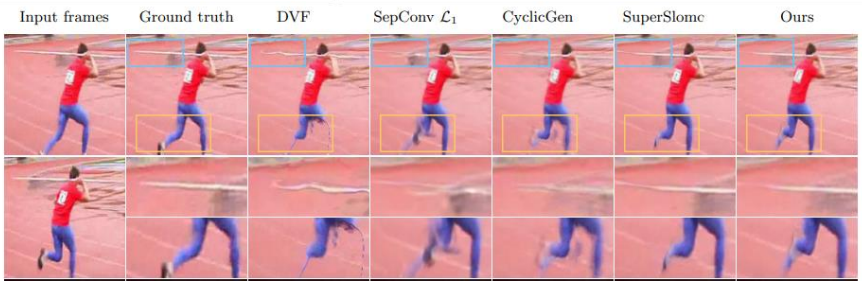


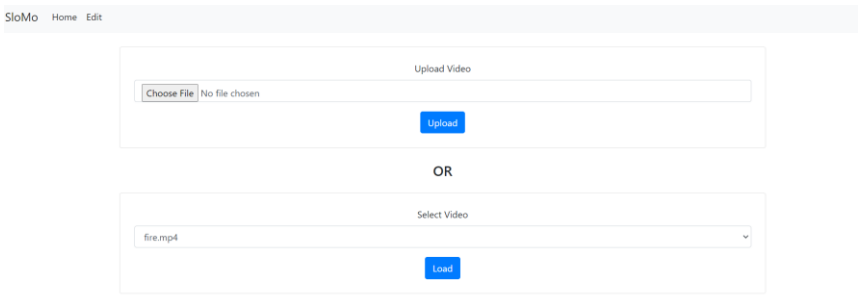
Fig. 21. Comparison of interpolated frames with state-of-art methods

With proper frame order, video reconstruction using FFmpeg can be done with no problems at all. The video is not formatted appropriately because some frames generated using this are not aligned equally to the size of the video frames. To see the movie, you'll need a specialised application such as VLC.

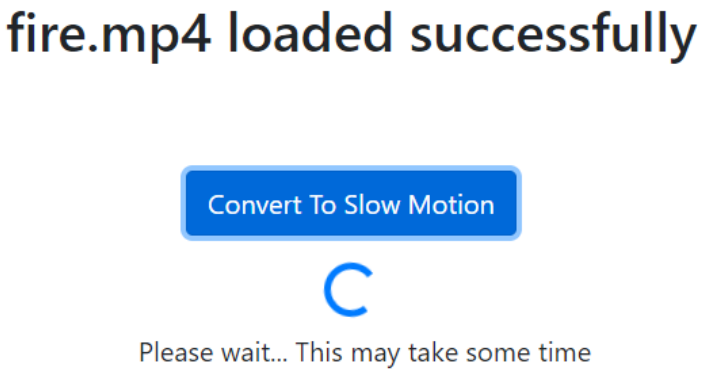
The movie created using this method has excellent slow-motion video quality since the data used to make it has been validated on the model beforehand, resulting in the highest degree of accuracy. The slow-motion module is tested on a variety of films with less than 220 frames in duration. Slow-motion video creation takes time; on average, it takes 5 minutes to convert a video to slow-motion. All of the videos produced acceptable results, including slow motion that could be understood and high quality.

The result of using the flask app .'s Steps and outputs for converting a video to slow motion are outlined below.:

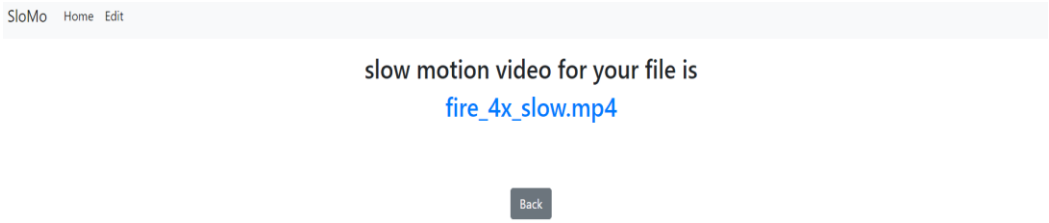
Step 1. load/upload a video.



Step 2. Convert the video to slow-motion and wait until conversion is done.

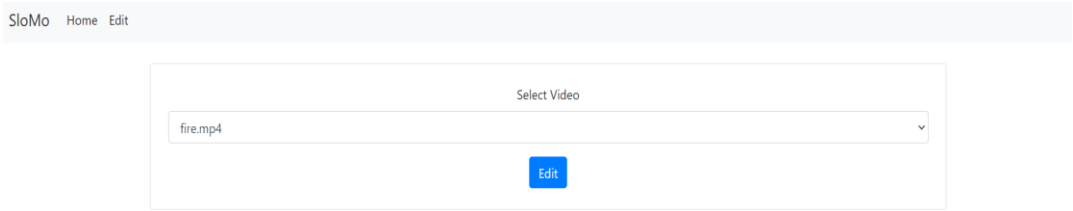


Step 3. Video is converted and the user can download the video.

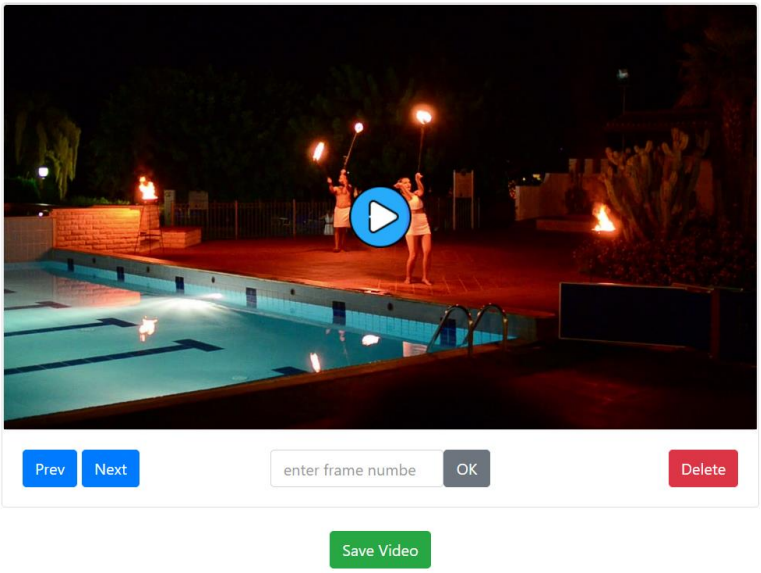


And To edit the video the following steps are needed to be followed.

Step 1. Go to the edit page and load the video.



Step 2. Navigate through the video and delete any irregular frames and save the video finally.



The Flask app is tested by using the Postman application in the windows. All the html pages are quality tested and improved through the reviews of several users. The application is unit tested unit testing facilities of python.

5.CONCLUSION

Two optical flows with pixel-wise weights are predicted from an unknown intermediate frame to the preceding and following frames. Warping the frames into the intermediate time step is done with flows. To create the intermediate frame, the warped frames are linearly concatenated with weights. It has been demonstrated that our CNN outperforms state-of-the-art approaches across a wide range of datasets after training on 1500 high-quality films. A new way for fine-tuning frame interpolation methods for each individual frame has also been created by our team. We've demonstrated that when combined with our approach, it yields better quantitative and graphical outcomes. It was possible to create a slow motion module using this approach, and then reconstruct the movie from interpolated frames using a recursive frame creation process. We recorded the recursively generated video frames. In addition, we created a web application that utilises the system's slow motion module so that users can interact with the system. We have introduced a video editing option that allows the user to remove undesired frames.

The goal of this project is to produce slow-motion footage. The bmp files generated by the interpolation network are saved in a temporary folder and can be accessed later. Using FFmpeg, the generated frames will be turned into a time-lapse video. These videos can be utilised to fully enjoy the scene and detect nuances that are often obscured by motion.

7.REFERENCES

[1] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In ICCV, 2017.
[2] "Hindu Kush", *SpringerReference*, Berlin/Heidelberg: Springer-Verlag, 2011.
[3] Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92(1), 1–31 (2011)
[4] Niklaus, S., Liu, F.: Context-aware synthesis for video frame interpolation. In: Conference on Computer Vision and Pattern Recognition. pp. 1701–1710 (2018)
[5] Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: International Conference on Computer Vision. pp. 4463–4471 (2017)
[6] Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9000–9008 (2018)
[7] Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: International Conference on Computer Vision. pp. 261–270 (2017)
[8] Liu, Y.L., Liao, Y.T., Lin, Y.Y., Chuang, Y.Y.: Deep video frame interpolation using cyclic frame generation. In: AAAI Conference on Artificial Intelligence (2019)
[9] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[10] Huang, Z., Zhang, T., Heng, W., Shi, B., & Zhou, S. (2020). RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation (Version 11). arXiv.
[11] Yapıcı, M. M., Tekerek, A., & Topaloğlu, N. (2019). Derin Öğrenme Araştırma Alanlarının Literatür Taraması. In Gazi Journal of Engineering Sciences (Vol. 5, Issue 3, pp. 188–215). Gazi Publishing.
[12] Parihar, A. S., Varshney, D., Pandya, K., & Aggarwal, A. (2021). A comprehensive survey on video frame interpolation techniques. The Visual Computer.