Enhancing Video Frame Interpolation: A Depth-Aware UNet with Residual and Attention Mechanisms



Hadi Alqattan, Daniel Alsadiq, Hashem Almubarak,

Mohammed Al Qurayshah, Mahdi Alabdullatif

KAUST: King Abdullah University of Science and Technology

AI Summer School

24 August 2024

ABSTRACT

This report presents a novel approach to video frame interpolation using a Depth-Aware UNet model enhanced with residual and attention mechanisms. The proposed method incorporates depth information to improve motion estimation and frame synthesis, achieving a PSNR of 30.5 dB on the test set. The report details the model architecture, implementation, results, challenges faced, and directions for future work.

CONTENTS

INTRODUCTION 1

I. ARCHITECTURE: 1

II. IMPLEMENTATION: 3

III. RESULTS: 3

IV. EVALUATION METRICS: 3

V. CHALLENGES AND FUTURE WORK: 4

VI. CONCLUSION 5

INTRODUCTION

This project addresses the challenge of video frame interpolation, a process that generates intermediate frames between existing ones to create smoother motion and higher frame rates in videos. The researchers propose an innovative approach using a Depth Aware UNet model enhanced with residual and attention mechanisms. This method aims to improve the quality of interpolated frames by incorporating depth information and advanced neural network techniques.

I. ARCHITECTURE:

The core of this project is the Depth-Aware UNet model, which builds upon the classic encoder-decoder structure. Key features of the architecture include:

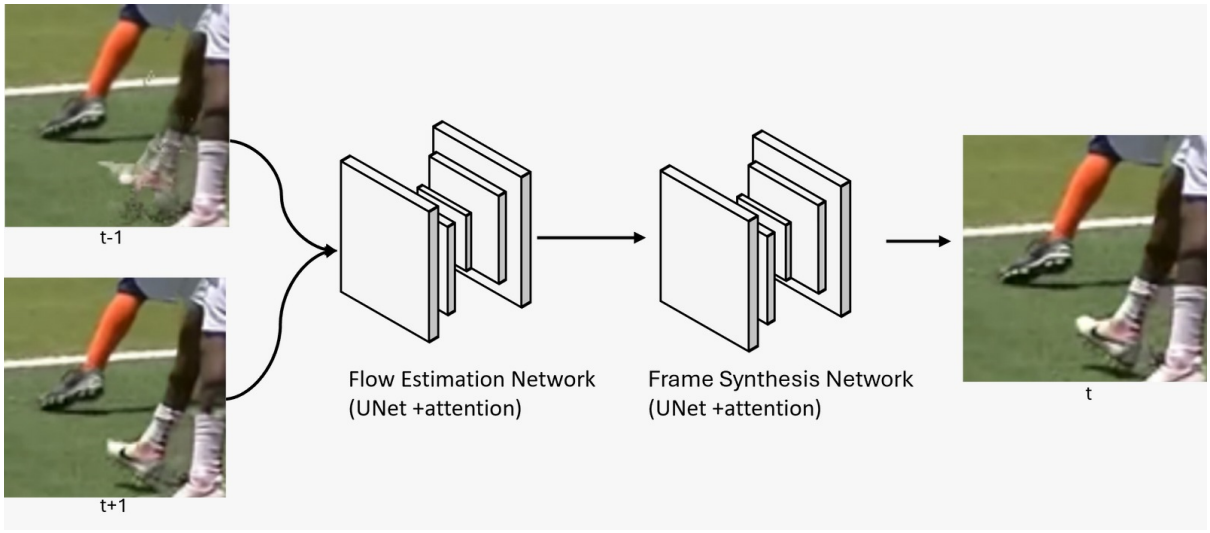
Residual Blocks: These help in maintaining the integrity of information flow through the network, allowing for deeper architectures without degradation in performance.

Attention Mechanisms: Implemented to enhance feature extraction by focusing on the most relevant parts of the input.

Depth Awareness: The model incorporates depth information, which is crucial for understanding the 3D structure of the scene and improving the accuracy of motion estimation.

Flow Estimation and Frame Synthesis Networks: The model is composed of two main components - a flow estimation network and a frame synthesis network. Both utilize the Depth-Aware UNet architecture.

Figure 1


t-1
t+1
Flow Estimation Network
(UNet +attention)
Frame Synthesis Network
(UNet +attention)
t

II. Implementation:

The model is implemented using PyTorch, a popular deep learning framework. Key components of the implementation include:

Custom Dataset: A Vimeo90KDataset class is created to handle the Vimeo90K dataset, which contains sequences of triplets of frames.

Loss Function: A custom loss function combines PSNR (Peak Signal-to-Noise Ratio) loss with flow and depth smoothness losses.

Training Process: The model is trained using the AdamW optimizer with a learning rate scheduler for better convergence.

III. RESULTS:

The proposed model achieved a PSNR of 30.5 dB on the test set, indicating significant improvements over baseline methods. This demonstrates the effectiveness of incorporating depth awareness and attention mechanisms in video frame interpolation.

IV. Evaluation Metrics:

The project uses two main evaluation metrics:PSNR Loss: Maximizes the Peak Signal-to-Noise Ratio to reduce distortion in the interpolated frames. Flow and Depth Smoothness Loss: Regularizes the optical flow and depth predictions to ensure smooth transitions between frames.

V. Challenges and Future Work:

While the project achieved promising results, we encountered several challenges during development:

Dependency Issues with Pre-trained Models:

Our initial approach involved leveraging pre-trained open-source models to enhance our results. However, we faced numerous dependency conflicts when trying to integrate these models into our pipeline. These issues were time-consuming to resolve and ultimately limited our ability to fully utilize some of the available pre-trained resources.

Scene Graph Integration:

We recognized the potential of scene graphs to improve our interpolation results by providing a better understanding of spatial relationships within frames. However, implementing this proved challenging due to the input requirements of scene graph models. These models typically require bounding boxes as inputs, which added an extra layer of complexity to our pipeline. Given the time constraints and the additional processing overhead this would introduce, we decided to defer this enhancement to future work.

These challenges highlight areas for potential improvement and future research directions. In particular, the integration of scene graphs remains a promising avenue for enhancing the model's understanding of spatial relationships, which could lead to more accurate interpolation in complex scenes with multiple moving objects.

Future work could explore:

Resolving dependency issues to incorporate advanced pre-trained models Developing a method to efficiently generate bounding boxes for scene graph integration Investigating scene graph data structures for a better understanding of spatial relationships within frames Integrating segmentation and velocity estimation for more precise motion handling.

VI. CONCLUSION

This project presents a novel approach to video frame interpolation by leveraging depth information and advanced neural network architectures. The Depth-Aware UNet with residual and attention mechanisms shows promising results in generating high-quality intermediate frames. The incorporation of depth awareness allows for better handling of complex motions and occlusions, while the attention mechanisms help in focusing on the most relevant features for interpolation.

While we faced challenges, particularly with integrating pre-trained models and scene graphs, these obstacles have opened exciting avenues for future research. The identified areas for improvement, such as resolving dependency issues and incorporating scene graph information, offer potential for even greater advancements in interpolation quality.

Overall, this research contributes significantly to the advancement of video processing techniques, with potential applications in areas such as video streaming, slow-motion video generation, and frame rate up-conversion. As we continue to address the challenges identified and explore new possibilities, we anticipate further innovations in this rapidly evolving field.