Deep Consecutive Attention Network for Video Super-Resolution

Mr. Talha Saleem

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Electrical Engineering
Department of Electrical Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2022
Copyright of Chulalongkorn University

โครงข่ายแบบเน้นความสนใจต่อเนื่องเชิงลึกสำหรับวีดิทัศน์ความละเอียดสูงยวดยิ่ง

นายทาลฮา ซาลีม

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	Deep Consecutive Attention Network for Video Super-Resolution
By	Mr. Talha Saleem
Field of Study	Electrical Engineering
Thesis Advisor	Associate Professor SUPAVADEE ARAMVITH, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

.... Chairman
(Assistant Professor Thavida Maneewarn, Ph.D.)
.... Thesis Advisor
(Associate Professor SUPAVADEE ARAMVITH, Ph.D.)
.... Examiner
(Assistant Professor SUREE PUMRIN, Ph.D.)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ทาลฮา ซาลีม : โครงข่ายแบบเน้นความสนใจต่อเนื่องเชิงลึกสำหรับวีดิทัศน์ความละเอียดสูงยวดยิ่ง. (Deep Consecutive Attention Network for Video Super-Resolution) อ.ที่ปรึกษาหลัก : รศ. ดร.สุภาวดี อร่ามวิทย์

งานด้านวีดิทัศน์ กระบวนการแสดงการเคลื่อนไหวช้าได้รับความสนใจมากในการสร้างวีดิทัศน์ความละเอียดสูงยวดยิ่ง กระบวนการสร้างวีดิทัศน์ความละเอียดสูงเคลื่อนไหวช้าจากเฟรมที่มีความละเอียดต่ำนี้ แบ่งออกเป็น 2 ส่วน ประกอบไปด้วยการสร้างวีดิทัศน์ความละเอียดสูงยวดยิ่งและการประมาณค่าเฟรมวีดิทัศน์ อย่างไรก็ดี แนวทางการประมาณค่าดังกล่าวนั้นไม่ประสบผลสำเร็จในการสกัดคุณลักษณะที่สนใจระดับต่ำ เพื่อให้ได้รับประโยชน์สูงสุดจากคุณสมบัติของความสัมพันธ์ของปริภูมิเชิงเวลา ในการแก้ไขปัญหาดังกล่าว เรานำเสนอวิธีการภายใต้การใช้โครงข่ายแบบเน้นความสนใจต่อเนื่องเชิงลึกสำหรับวีดิทัศน์ความละเอียดสูงยวดยิ่ง จากการใช้กลไกเน้นความสนใจหลายหัวและโมดูลเน้นความสนใจของคุณลักษณะเชิงเวลา เพื่อให้การคาดเดาการประมาณค่าคุณลักษณะเฟรมได้ดีขึ้น โมดูลคอนแอลเอสดีเอ็ม (ConvLSTM) ที่เปลี่ยนรูปได้แบบสองทิศทาง และจัดกับข้อมูลจากกลไกเน้นความสนใจหลายหัวและกลุ่มคุณลักษณะเชิงเวลาเพื่อเพิ่มคุณภาพของเฟรมวีดิทัศน์ วิธีการนี้สังเคราะห์เฟรมวีดิทัศน์ที่มีความละเอียดสูงขึ้นมาจากเฟรมวีดิทัศน์ที่มีความละเอียดต่ำ ผลการทดลองแสดงให้เห็นว่าวิธีการโครงข่ายแบบเน้นความสนใจต่อเนื่องเชิงลึกที่นำเสนอมีประสิทธิภาพในเชิงค่าพีเอสเอ็นอาร์สูงกว่า 0.27 เดซิเบลและ 0.31 เดซิเบลโดยเฉลี่ยสำหรับชุดข้อมูลทดสอบ Vid4 และ SPMC ตามลำดับ เมื่อเปรียบเทียบกับวิธีการทันสมัยที่เป็นพื้นฐานในปัจจุบัน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาขาวิชา	วิศวกรรมไฟฟ้า	ลายมือชื่อนิสิต ....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

6370396521 : MAJOR ELECTRICAL ENGINEERING
KEYWORD: Video Super-Resolution, Multi-head Attention, Attentive Feature Temporal Interpolation, Convolutional neural network

Talha Saleem : Deep Consecutive Attention Network for Video Super-Resolution. Advisor: Assoc. Prof. SUPAVADEE ARAMVITH, Ph.D.

In the video application, slow motion is visually attractive and gets more attention in video super resolution. To generate the high-resolution (HR) slow motion video frames from the low-resolution (LR) frames, two sub-tasks are required, including video super-resolution (VSR) and video frame interpolation (VFI). However, the interpolation approach is not successful to extract low level feature attention to get the maximum advantage from the property of space-time relation. To this extent, we propose a deep consecutive attention network-based method. The multi-head attention and an attentive temporal feature module are designed to achieve better prediction of interpolation feature frame. Bi-directional deformable ConvLSTM module aggregates and aligns with the information from the multi-head attention and temporal feature block to improve the quality of video frames. This method synthesizes the HR video frames from LR video frames. The experimental results in terms of PSNR show the proposed method of deep consecutive attention outperforms 0.27 dB and 0.31 dB for Vid4 and SPMC datasets respectively, in average of PSNR compared to state-of-the-art baseline method.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study:	Electrical Engineering	Student's Signature ....
Academic Year:	2022	Advisor's Signature ....

ACKNOWLEDGEMENTS

My warmest gratitude to my advisor, Associate Professor Dr. SUPAVADEE ARAMVITH for her guidance and encouragement. I have greatly benefited from her wealth of knowledge. I am immensely thankful that she took me on as her master’s degree student and continued with faith in me throughout the degree.

Thank you to Assistant Professor Dr. THAVIDA MANEEWARN and Assistant Professor Dr. SUREE PMRIN for being the committee members. Their thoughtful and fruitful feedback has been important to me in the completion of this research.

Additionally, I am very thankful to my seniors, Watchara Ruangsang, Sovann Chen, and Amir Hajian, to give me their time for valuable suggestions. They always encourage and support me in my research work.

I gratefully acknowledge the financial support from Chulalongkorn University's Graduate Scholarship Programme for ASEAN or Non-ASEAN Countries.

Finally, I am thankful my parents who supported and motivated me in my life. I am so grateful to my VTRG lab-mates and friends for giving me continuous support.

Talha Saleem

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

TABLE OF CONTENTS

Page

ABSTRACT (THAI) iii

ABSTRACT (ENGLISH)...iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS...vi

LIST OF TABLES..ix

LIST OF FIGURES x

CHAPTER 1 INTRODUCTION 1

1.1 Motivation and Research Problem...1

1.2 Objectives 8

1.3 Scope of Research...8

1.4 Expected Output 8

CHAPTER 2 RELATED WORKS...9

2.1 Video Frame Interpolation (VFI)..9

2.1.1 Phase-based approach...11

2.1.2 Flow-based approach...12

a. Optical Flow Based Interpolation...13

b. Motion Compensated Interpolation..14

2.1.3 Kernel-based approach..15

2.1.4 Deformable-convolution-based approach 17

2.1.5 Visual artifacts and occlusion...18

2.2 Video Super-Resolution (VSR) 18

2.2.1 Methods with Alignment...20

a. Deep-DE 20

b. VSRnet 20

c. VESPCN..21

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

d. DRVSR 21
e. FRVSR 22
f. SOFVSR 22
g. TOFlow 23
2.2.2 Methods with Deformable convolution 23
a. EDVR 24
b. DNLN 24
c. TDAN 25
d. D3Dnet 25
e. VESR-Net 25
2.2.3 Methods without alignment 26
a) 2D convolution methods 26
b) 3D convolution methods 27
2.2.4 Recurrent back-projection network 27
2.2.5 Convolutional LSTMs 27
2.3 Channel Attention (CA) 28
2.3.1 Channel attention module 28
2.3.2 Scaled Dot-Product Attention 29
2.3.3 Multi-Head Attention 30
2.4 Space-Time Video Super-Resolution (STVSR) 31
CHAPTER 3 PROPOSED METHOD 33
3.1 Overall Framework 33
3.2 Deep consecutive attention model 34
3.2.1 The early stage of feature extraction 34
3.2.2 Multi-head Attention 34
3.2.3 Attentive feature temporal interpolation 38
3.3 Bidirectional deformable ConvLSTM 40
3.4 Frame Reconstruction 41
3.4.1 Pixel Shuffle 41

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER 4 EXPERIMENTAL RESULTS AND DISCUSSIONS 43
- 4.1 Experimental Setup 43
- 4.2 Evaluation metrics 43
  - 4.2.1 Peak signal-to-noise ratio (PSNR) 43
  - 4.2.2 Structural similarity index measure (SSIM) 43
- 4.3 Datasets 44
- 4.4 Evaluation 46

CHAPTER 5 CONCLUSION 54
- 5.1 Conclusion 54
- 5.2 Future Work 54

REFERENCES 55

VITA 62

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF TABLES

Page

Table 1: Comparison of the proposed method on Vid4 dataset 53

Table 2: Comparison of the proposed method on SPMC dataset 53

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

LIST OF FIGURES

Page

Figure 1: Video frame of same scene with different resolutions 2

Figure 2: General process for video frame 2

Figure 3: Distorted shape of ball in fast motion 4

Figure 4: Image of occlusion and motion ambiguity challenges 5

Figure 5 Pipeline of VSR tasks..7

Figure 6: A taxonomy of video frame interpolation approaches 9

Figure 7: Depth-aware VFI network..10

Figure 8: PhaseNet block by Meyer..12

Figure 9: Context-aware Frame Synthesis 12

Figure 10: Optical flow between pixels 13

Figure 11: Occlusion and Optical flow between pixels 13

Figure 12: Framework of interpolation based on optical flow 14

Figure 13: Full search motion 14

Figure 14: Context extraction network 16

Figure 15: Residual block 17

Figure 16: Deformable convolution approach 18

Figure 17: Basic concept of VSR tasks...19

Figure 18: A taxonomy for VSR methods 19

Figure 19: The architecture of Deep-DE 20

Figure 20: VSRnet architecture 20

Figure 21: The network architecture of VESPCN 21

Figure 22: Motion estimation and compensation..21

Figure 23: Architecture of DRVSR 22

Figure 24: Architecture of FRVSR..22

Figure 25: Architecture of SOFVSR..22

Figure 26: Architecture of TOFlow 23

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Figure 27: Deformable convolution 23

Figure 28: The network architecture of EDVR 24

Figure 29: Overview of Enhanced Deformable Video Restoration (EDVR) 24

Figure 30: Architecture of DNLN 24

Figure 31: Overview of Temporally Deformable Alignment Network (TDAN) 25

Figure 32: Architecture of D3Dnet 25

Figure 33: VESRNet architecture 26

Figure 34: The architecture of the generator in VSRResFeatGAN 26

Figure 35: The architecture of FFCVSR 26

Figure 36: Overview of Dynamic Up-sampling Filters (DUF) 27

Figure 37: Overview of RBPN 27

Figure 38: Channel Attention Module 28

Figure 39: Attention with Scaled Dot-Product 29

Figure 40: Attention with Multi-Head 31

Figure 41: Overview Zooming Slow-Mo 32

Figure 42: The framework of Deep Consecutive Attention Network for VSR (DCAN) 33

Figure 43: Query, Key and Value 35

Figure 44: Calculation of attention Score 36

Figure 45: Scaled Dot-Product Attention 36

Figure 46: Multi-head attention 37

Figure 47: Attentive feature temporal interpolation module 39

Figure 48: Deformable ConvLSTM 40

Figure 49: Pixel Shuffle 42

Figure 50: Example images from Vimeo 90K dataset 44

Figure 51: Example images from SPMC dataset 45

Figure 52: Example images from Vid4 dataset 46

Figure 53: Visual comparisons of frames of “LDVTG_009” from SPMC dataset on scale ×4 47

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Figure 54: Visual comparisons of frames of “NYVTG_006” from SPMC dataset on scale ×4 47

Figure 55: Visual comparisons of frames of “Veni3_011” from SPMC dataset on scale ×4 48

Figure 56: Visual comparisons of frames of “hdclub_001” from SPMC dataset on scale ×4 48

Figure 57: Visual comparisons of frames of “car05” from SPMC dataset on scale ×4 49

Figure 58: Visual comparisons of frames of “jvc_004” from SPMC dataset on scale ×4 49

Figure 59: Visual comparisons of frames of “calendar” from Vid4 dataset on scale ×4 50

Figure 60: Visual comparisons of frames of “walk” from Vid4 dataset on scale ×4 . 50

Figure 61: Visual comparisons of frames of “city” from Vid4 dataset on scale ×4 ... 50

Figure 62 Visual comparison of video frames of “city” 51

Figure 63 Visual comparison of video frames of “foliage” 52

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

CHAPTER 1

INTRODUCTION

1.1 Motivation and Research Problem

There are numerous important moments in the daily routine of our life that everyone wants to record with a camera, i.e., baby first time strolling, some tricky skateboard moves, playing with loving pets. It is conceivable to have 240 frame-per-second (fps) video with a smartphone, high-speed professional cameras are however expected a high frame rate. Furthermore, millions of images and videos are published daily. Users can access these multimedia streams to learn more about topics that interest them or share special images with their loved ones. People appear to want high resolution more and more, mainly because they want a realistic and vivid visual experience. This is another aspect of the market-driven production of Ultra High-Definition TV by manufacturing businesses today [1]. However, the resources you often access are low-resolution because they have either been compressed to fit within uploading constraints or are constrained by the capabilities of devices. The use of high-resolution displays in homes and mobile devices is increasingly widespread. Therefore, everyone demands for methods to convert LR images and videos into HD versions [2]. Reconstructing the HR version from the LR version is known as an image or multi-frame super-resolution. A video is a collection of images shown one after the other at a set frame rate. Video super-resolution is supposedly more flexible than single-image super-resolution because it can choose from various useful pieces of information to get superior results. The same scene of a video frame sequence is shown in Figure 1 with different resolutions. Consequently, numerous multimedia applications are expected to embrace it, including surveillance cameras [3], video streaming [4], high-definition television [5], video compression [6], [7], remote sensing [8], and video conferencing [9]. Many life events, like opening a bottle of champagne or seeing lightning, take place quickly and are challenging to watch in real time. A high frame rate camera can capture these moments effectively, and slow-motion resulting video can be viewed. Slow-motion videos are becoming more and more common since modern smartphones can capture high frame rate video.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Low Resolution 112×64

High Resolution 448×256

Figure 1: Video frame of same scene with different resolutions

However, due to their constrained bandwidth, these cameras boost their frame rate at the expense of their spatial resolution. Moreover, a significant number of the events we might like to have with slow-motion are unpredictable, and these events are recorded with the frame rates having defined standard. It is impractical to record videos having a high rate of because it requires considerable memory and power for smart phones. Hence it is phenomenal to generate video with slow-motion having high quality from the recorded ones. Making a video in slow-motion from high-quality recorded ones is incredibly interesting. These converted videos with higher frame rates benefit from smoothness due to video interpolation [10].

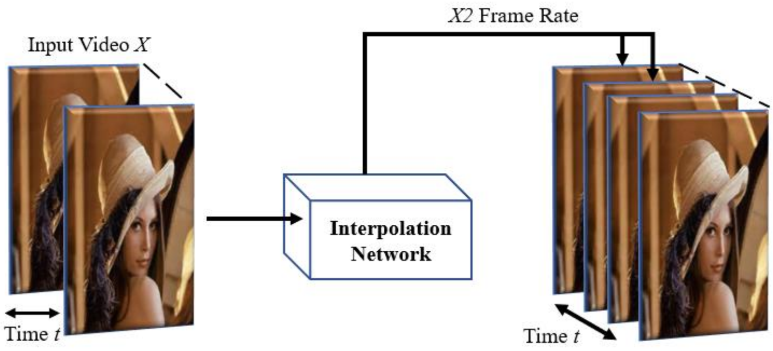


Figure 2: General process for video frame

Other intriguing new uses include learning optical flow through the analysis of unlabeled videos utilizing a supervisory signal. Usually, a camera has limited spatial resolution and temporal resolution. The sensor’s spatial density in the camera and the blur of these sensors describes the spatial resolution. These reasons regulate the miniature size of spatial features detected visually. The video camera’s frame rate and time describe temporal resolution. This normalizes the utmost speed of active events reflected in a video sequence. Continuous and streaming visual data is typically recorded in the video with discrete consecutive frames. While the high-fidelity video is luxurious, it primarily accumulates at low-resolution (LR) and low-frame rates (LFR). The effort on spatiotemporal video super-resolution (VSR) has been recently established to blend temporal and spatial interpolation in a cohesive framework. It is expected to have a high-level spatiotemporal resolution of video sequences. Videos or images having large sizes can now be delivered in less time after the development of 5th-generation mobile communication technology. In the meantime, video super-resolution is garnering greater attention due to the increasing recognition of high-definition (HD) and ultra-high-definition (UHD) display. Video is one of the most dominant forms of multimedia, so enhancing low-resolution videos has become crucial. While VSR algorithms deal with many consecutive images or frames at same time to take advantage of relationships between frames to resolve the targeted frame, image SR techniques typically handle one image at one time. VSR can manage frame by image SR algorithms and is seen as an increase of image SR. Still, the SR performance is always not excellent due to the prospect of introducing artifacts and jams, which results in unwelcome temporal incoherence within frames, [11].

Instant dynamical events that occur faster than the frame rate of cameras is not seen or captured improperly in recording of video sequences. This matter is often seen in action-packed sports, such as tennis, wrestling, and hockey, where it is hard to discern the motion or the actions of the ball or puck due to its rapid movement. In video sequences, two common visual effects are brought on by swift motion. The exposure duration of the camera is the cause of motion blur, and the temporal sampling presented by the frame rate of camera because of the other effect aliasing in motion.

(i) *Motion Blur*: The camera creates frame by combining the scenes during the exposure time. Objects having fast motion therefore cause a visible blur beside their path, frequently leading to deformed or unrecognizably shaped things. This impact intensifies with speed, notably if the object's trajectory deviates from a straight line. Without distinguishing motionless and dynamic scene features or approximating their motions, spatial aberrations like blur in motion can be handled by increasing the temporal resolution utilizing information from different video sequences [12]. Contrary to spatial SR, motion-deblurring in temporal SR requires minimal input cameras (video sequences) [13].

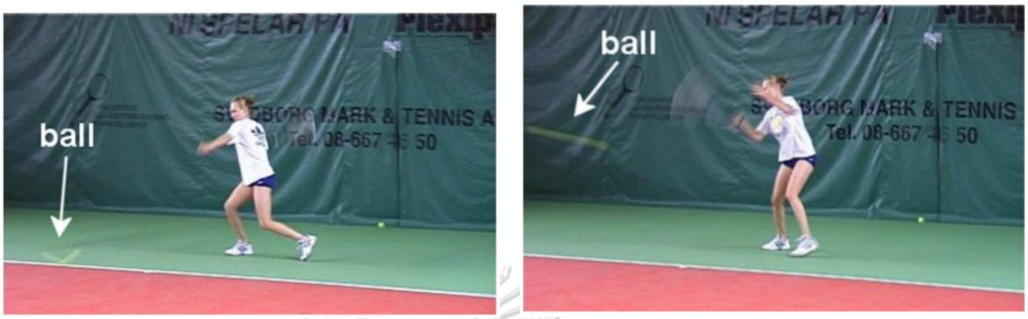


Figure 3: Distorted shape of ball in fast motion

(ii) *Motion Aliasing*: Incorrect visual effects taken on by aliasing in point in time are a very significant concern in sequences of video with quickly changing situations. When a fast-moving object generates a trajectory with frequencies greater than the camera's frame rate, this is known as motion aliasing. The temporal resolution having high frequencies are "folded" into the temporal resolution having low frequencies. A distorted or inaccurate trajectory of the object with motion is visible. The well-recognized optical effect known as "wagon wheel effect" illustrates motion-based aliasing. When a wheel is quickly spinning, over a particular motion, it will seem revolving in the "wrong" way [14].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Temporal interpolations with complexity are used to increase the frame rate. The aliasing and blur because of motion cannot be improved by performing such sequences of video in "slow motion." This is so that the missing details of extremely quick dynamic occurrences can be retrieved from the information in one sequence of videos. A huge number of frames are extracted per second while creating slow-motion videos. If we don't record enough frames, the fast video will look rough and unwatchable unless we use sophisticated AI techniques to expect the additional frames by applying deep learning procedures to convert appealing, 240 fps slow-motion from 30 fps video. The framework of AI chooses two different frames and then builds intermediate motion by following the sequence of objects from one frame to the other. Although it doesn't provide outcomes quite as accurate as a human intellect would, it comes near. Before the technique can prosper commercially, it needs to be improved. The extreme temporal resolution has been missed because of too much time being subsampled and blurred [15] shown in Figure 4.

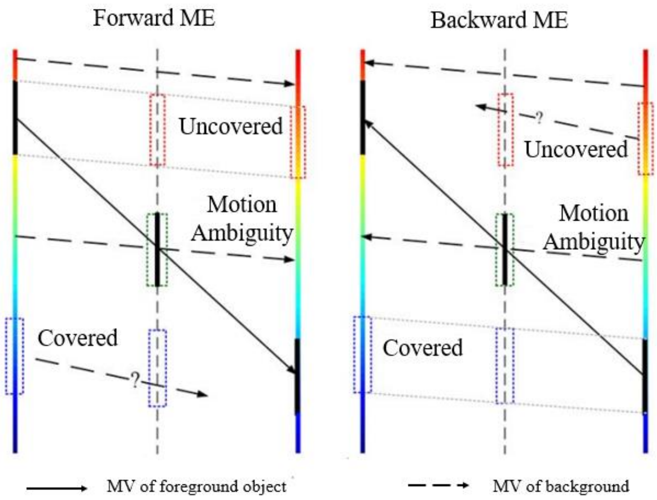


Figure 4: Image of occlusion and motion ambiguity challenges

On the other hand, numerous sequences offer more illustrations of the dynamic scene. While none of the specific sequences has adequate visual information, combining the data from every sequence facilitates the creation of a high-resolution video sequence that accurately depicts dynamic occurrences. Thus, even though the wagon wheel appears erroneously in all the input sequences, a reconstructed HR sequence will appear as the appropriate motion. [15].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Regardless of having extremely different attributes, the spatial dimensions and temporal dimensions are associated. This introduces spatial and temporal visual tradeoffs particular to spatiotemporal SR and is not present in more conventional spatial SR. For example, the same input sequences can create output sequences with various space-time resolutions. Typically, a high increase in the spatial resolution must be offset by a substantial increase in temporal resolution. Additionally, blending input images with various spatial resolutions is not beneficial in classic image-based SR because a HR image will merge the data in a LR image. Information from several camera kinds and space-time resolutions may be complementary. To create a better video sequence with a more spatial and more temporal resolution, data can be a combine from high-quality cameras—which have good spatial resolution but have very less "temporal resolution"—with information from traditional video cameras. Motion interpolation occasionally accompanies visual anomalies in the screen, such as a tiny tear or glitch that appears briefly. The innovation's impact is most obvious when it suddenly appears during a rapid camera pan. The number of artifacts in contemporary commercial displays has decreased over time due to the advancement of related technology but has not eliminated it. It is crucial to research and fill in the significant space between frames. The soap opera effect, however, destroys the theatrical appearance of the film works by making it appear as though the viewer is either on set or seeing the background elements [10], [15]. As a result, almost all manufacturers have sought to reduce the feature's functionality or impact quality. The main issue of a typical solution in problem setups that generate several intermediate frames is not only to approximate appropriate motion among successive images but also to monitor occlusion to avoid excessive artifacts across motion limitations in the interpolated output. Motion interpolation is widely accepted despite its flaws because it improves visual clarity by dropping motion uncertainties by camera shakes and defective cameras. Additionally, it makes it easier to develop computer games with a better frame rate for a further convincing experience, although the additional input lag can be an involuntary side effect. The key differences between a frequently captured high frame rate and an inserted high frame rate are that the latter is reliant on none of those as overhead mentioned flaws, contains increasingly accurate image data and demands extra memory and transmission speed for the reason that frames are not formed in real-time.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

High-speed internet, efficient storage space opportunities, and high-proportion compression methodologies, like MPEG-1, MPEG-2, and MPEG-4, have all made it simpler to comprehend accessible video data. Therefore, there is a considerable requirement for the automated detection of semantically significant results for video reviews to aid in video utilization, handling, and indexing. Computer vision literature comprises many methodologies to programmed event-based recognition and outlining in action-packed sports applications. Most tactics, resulting in domain-specific methodologies, are developed for certain games, visual editing, or explicit scenarios. For instance, some limit the events to football games, while others limit them to baseball, soccer, or basketball. Some of them also mandate that cameras observe the events [10].

Because of the motion dynamics of high-resolution (HR) video with slow motion, it is visually evident and finely detailed. However, maintaining the perceptual quality and photo-realistic video sequence while converting LR and LFR videos into HR is a crucial challenge [16]. Shechtman *et al.* [14] recently developed the Space-Time VSR model, which enhanced the temporal resolution and spatial resolutions from the LR sequence. Additionally, the application of this STVSR model in real-world scenarios is constrained by its high processing cost and poor performance in video sequences with fast-varying motions and complex analytic forms. VFI and VSR are two machine learning tasks where deep learning-based techniques [17], [18] have shown promise in recent years. First, the missing intermediate LR frames from the sequences are interpolated with VFI, and after this HR frames are recreated with VSR. The spatial resolution and the time interpolation are connected in this instance. This method fails to fully use the space-time relation property because it cannot draw attention to low-level features.

SR methods established on deep learning are deeply researched because of the massive accomplishment of deep learning in many disciplines. Many deep neural network VSR methods have been created. They frequently use LR and HR sequences to input the network for alignment within the frame, extraction of features, and their fusion. They create HR sequences for the corresponding LR sequences. Many VSR methods have a pipeline consisting of a module of alignment, a feature extraction and a reconstruction [11], as shown in Figure 5.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

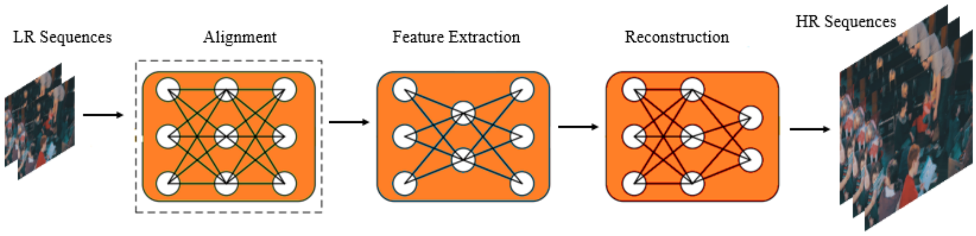


Figure 5 Pipeline of VSR tasks

Latest convolutional neural network (CNN) based techniques to create a unified framework [19], [20], [21] have been developed to solve these shortcomings. This means that within a deep learning network, the LR features are first retrieved from the LFR and then up-sampled in time and space. Although the unified end-to-end framework outperforms the previous models in terms of implementation, the state-of-the-art baseline method [19] cannot completely exploit mutual relations because of shortcomings in the outcomes of feature temporal interpolation. Due to this flaw, reconstruction errors accumulate in both temporal and spatial domains, leading to undesirable effects in the super-resolved results, such as blurring and aliasing.

The thesis is structured as follows: Chapter 2 explains the related works comprising of video frame interpolation (VFI), video super-resolution (VSR), channel attention (CA), Multi-head attention (MHA), and space-time video super-resolution (STVSR). The overall framework with a deep consecutive attention network is introduced in chapter 3 of the methodology. Experimental results of deep consecutive attention networks for video super-resolution have been demonstrated in chapter 4. Lastly, the conclusion of the research work is described in chapter 5.

1.2 Objectives

a. Develop an algorithm to generate high-resolution (HR) slow-motion video sequences.
b. Develop a deep consecutive multi-head channel attention network for video super-resolution (VSR).

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.3 Scope of Research

a. Measured the performance of the proposed video super-resolution (VSR) model by PSNR and SSIM at scale ×4.
b. Evaluate the performance of the proposed video super-resolution (VSR) algorithm with state-of-the-art methods in terms of objective and subjective scores.

1.4 Expected Output

a. Produce better quality slow-motion HR video output compared to previous models.
b. The algorithm can be operated in real-time scenarios.

CHAPTER 2

RELATED WORKS

This chapter goes through the related work on video frame interpolation (VFI), video super-resolution (VSR), Channel Attention (CA), Multi-head attention (MHA), and space-time video super-resolution (STVSR), respectively.

2.1 Video Frame Interpolation (VFI)

The key goal of the VFI is to generate smooth motion with the least amount of visual blur while maintaining the local information of objects with motion in the produced intermediate frames. By creating an intermediate frame from two neighboring original frames [22], VFI tries to improve temporal resolution. The picture sequence estimation problem serves as the foundation for conventional approaches. This method falls short in scenes with intricate image textures and quick movements. This VFI technique includes path-based and phase-based. There are three types of interpolation methods: flow-based, GAN-based, and CNN-based, as shown in Figure 6.

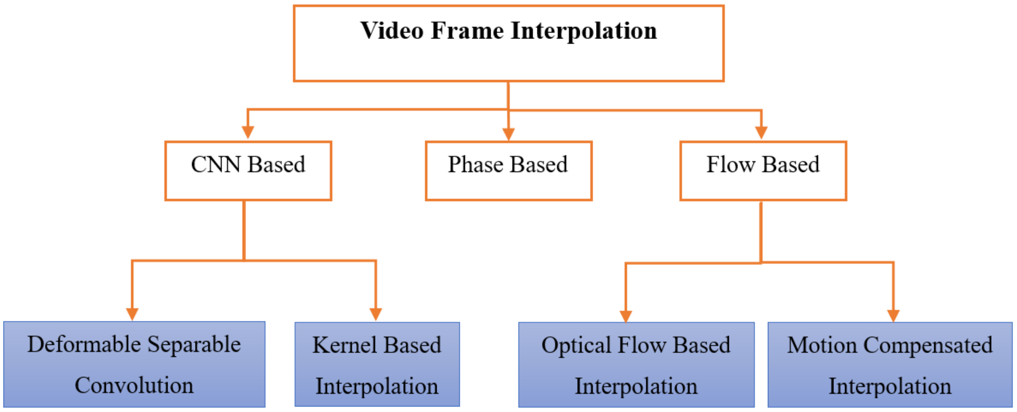

Video Frame Interpolation
CNN Based
Phase Based
Flow Based
Deformable Separable Convolution
Kernel Based Interpolation
Optical Flow Based Interpolation
Motion Compensated Interpolation

Figure 6: A taxonomy of video frame interpolation approaches

A self-supervised framework was introduced by Liu *et al.* [1] and automatically modifies the network to calculate improved optical flow and distort images to make an intermediate frame. Compared to traditional supervised techniques, substantial results are obtained. However, their approach fails the optical flow estimate method because it generates undesirable effects like ghosts and halo because of occlusion.

Liu *et al.* [23] introduce method to enhance optical flow estimation by the captivating closeness among input and mapped-back images. This method also used motion linearity to work with large-scale motion and rich texture problems. Still, the method does not exhibit improved occlusion and complex motion results. There are now practical approaches for processing occlusion. Tianfan *et al*. [24] formed a system with three sub-associations, where main network combines interpolated frames using estimated parameters.

In contrast, the networks calculate occlusion mask and optical flow. Jiang *et al.* [25] applied visibility maps that only blend the unoccluded picture element into the interpolated image to address occlusion analysis. In response to this issue, Bao *et al.* [26] work on depth awareness, shown in Figure 7. These methods outperform other techniques with glaringly superior outcomes. However, these interpolation techniques fall short when used with high-resolution movies beyond 4K and substantially record large-scale motion. Color and depth consistency can help to enhance the accuracy of estimated depth maps [27].

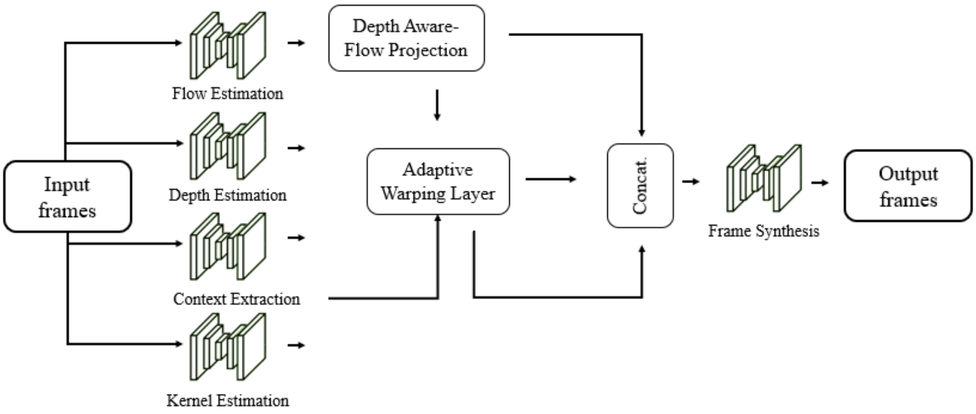


Figure 7: Depth-aware VFI network

Contextual data and optical flow were both used by Niklaus *et al.* [28] to create a context-aware network. This system is created from Gridnet [29], which take together warping and pixel blending in one stride, in contrast to conventional interpolation techniques. Despite the significant state-of-the-art performance, this approach cannot analyze HR frames of video due to the inherent network complexity's memory constraints. Meyer *et al.* [30] designed a convolution neural based network, this model is not well-known for its texture, to control huge motion successfully. This is illustrated in Figure 8. A network created by Niklaus *et al.* [31] estimates pixel-wise spatially adaptable kernels, including information about the optical flow between successive input frames with pixel warping. Their technology requires more computer resources when processing high-resolution video frames but offers cutting-edge performance for straightforward small-scale movements. Niklaus *et al.* [32] method substitutes the interpolation of 2-dimentional kernels frame with two distinguishable 1-dimentional kernels to resolve the high-level memory requirement. Their method cannot process video frames at a resolution of 4K or higher since it requires more computation than current approaches. A remainder learning method that combines a multi-level residual estimate section, which creates the synthesized frame and expected flow, this method was suggested by Amersfoort *et al.* [33]. Interpolated economic motion neural networks for resolution with HD have been proposed by Peleg *et al.* [34] and Vidanpathirana *et al.* [35]. They suggest a real-time temporally aligned frame output in a block-wise way on CNN platforms [36]. A hybrid network created by Ahn *et al.* [37] comprises of temporal interpolation and spatial interpolation sub-layers that gradually generate a intermediate frame with high-quality subject to large-scale motion and complicated structural alterations cutting-edge performance is achieved.

2.1.1 Phase-based approach

Frame interpolation having phase information was primarily applied by Meyer *et al.* [22]. This methodology was established on the concept that for each pixel's phase shift values it will deliver small motion information. Though, they could not get their system to work well for significant motion. The establishment of phase-based methods is the idea that the motion of certain signals can be signified as a phase shift. The system's aim is to create an intermediate image from the input of two neighboring images. Our network directly predicts the steerable pyramid decomposition values rather than the color pixel values. They suggested a multi-scale pyramid-level structure Meyer *et al.* [30] to propagate phase information.

CHULALONGKORN UNIVERSITY

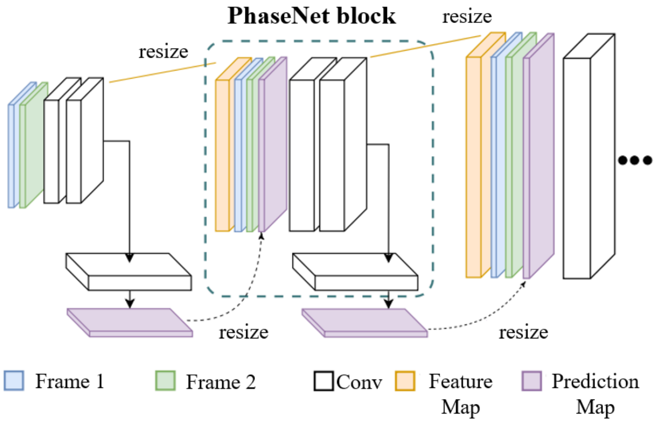


Figure 8: PhaseNet block by Meyer

2.1.2 Flow-based approach

Flow-based algorithms [25], [28], improve temporal resolution by calculating the optical flow between neighboring frames. To enhance the feature of the accompanying video, it is required to explicitly synthesize intermediate frames and identify the kind of flow among appropriate objects in successive frames.

Jiang *et al.* [25], model shown in Figure 8, this network learns to explain the presence of the two images taken as input in supplement to the motion models. The massive RGB color space makes it difficult to generate high-quality intermediate images. Figure 10 illustrates an approximation for optical flow for intermediate frame.

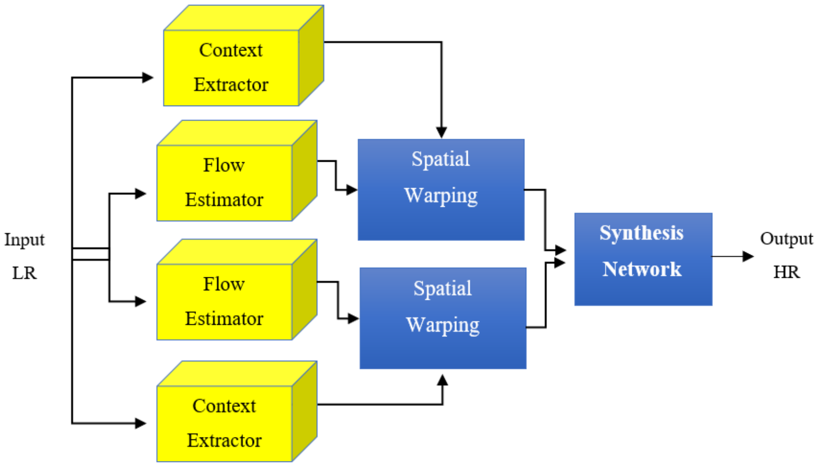


Figure 9: Context-aware Frame Synthesis

Niklaus *et al.* [28], this technique bends the frames taken as input and context maps o these inputs by optical flow. It uses forward warping that makes use of optical flow, specifically. This method produced holes in the distorted output, primarily occlusion, as shown in Figure 11.

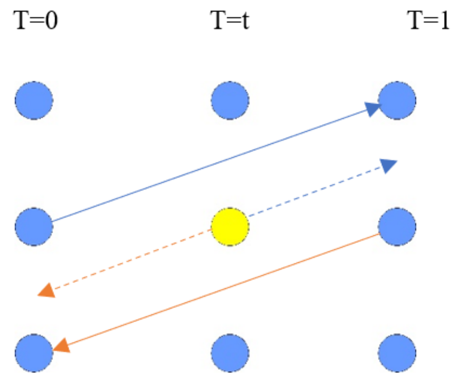


Figure 10: Optical flow between pixels

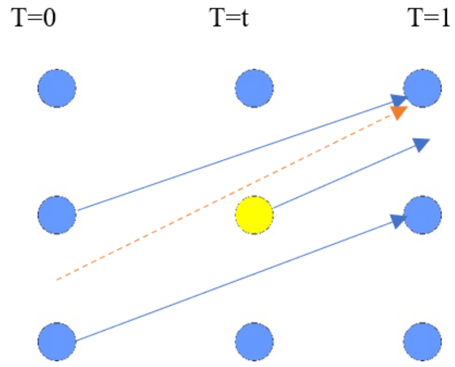


Figure 11: Occlusion and Optical flow between pixels

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

a. Optical Flow Based Interpolation

This technique delivers upscaling of frame rate by leveraging optical flow in bidirectional, which identifies information of motion among consecutive images and gathers dense pixel similarities. The visible motion of frames moving with bi-dimensional motion space is recommended by optical flow, and it can be examined as a trouble for the image interpolation area. Traditional methods utilized a variational version with an energy-saving method. Another research topic is improving the pixel synthesis stage, which involves merging the pixels of warped frames to produce an interpolated frame. Occlusion is handled concurrently with the aid of bidirectional information [38], [39], but is limited to pixel-by-pixel blending. As demonstrated in Figure 12, latest pixel synthesis method based on deep learning like Super SlowMo [25] and CtxSyn [28] .

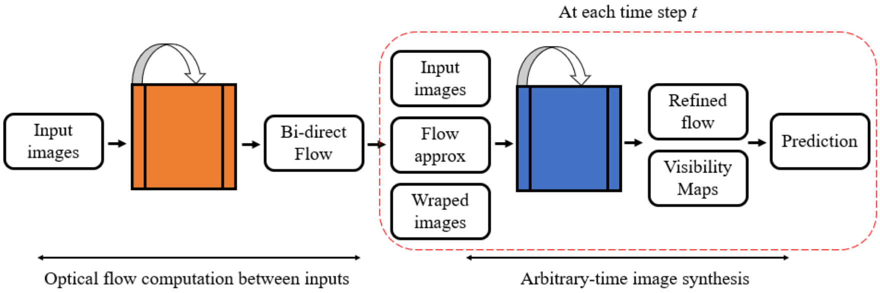


Figure 12: Framework of interpolation based on optical flow

Super SloMo [25], SepConv [32], and DAIN [26] are traditional techniques that have so far attained state of the art resolution of videos of the UCF101 [40] and Middlebury [38] standards. In contrast to earlier methods, it rebuilds information instead of just adjusting pixel information to decrease the blur generated by traditional video enhancement methods [41]. Implementation is evaluated using improved Vimeo datasets, SJTU Media datasets, and Ultra Video datasets with sufficient 4K image resolution.

b. Motion Compensated Interpolation

This method uses motion vectors to determine the transformation strategy from a given reference frame to the target frame. These techniques were developed to address the shortcomings of prior non-motion-compensated frame estimation techniques. It is a three-step sequential motion vector smoothing, motion estimation (ME), and motion-compensated (MCI) interpolation. The "velocity" vector of motion estimate is programmed to be calculated. Every pixel in the input frame, or the path, is taken by a picture in a time frame. Additionally, estimated motion determines constituent motion, and vectors spatially compensate each pixel halfway.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

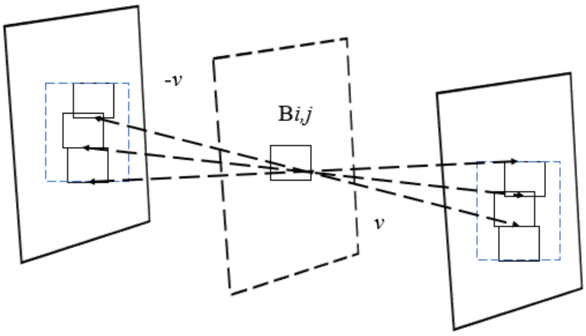


Figure 13: Full search motion

This work seeks to assess the accomplishment of modern up-conversion of frame rate with advanced MC techniques. Using a multi-hypothesis Bayesian FRUC model, Hongbin Liu *et al.* [42] unified a model to revise the optimization standard for forecasting the interpolated frame. As a replacement for optimal solution, the model builds a group of motion trajectory hypotheses using a set of "optimal" motion fields, as shown in Figure 13. After examining the numerous behaviors, dependencies, and interactions between different levels of consecutive video frames, Zhefei Yu *et al.* [43] proposed a self-correcting multilevel model that included a three-pixel block level, and sequence level. Each level implements constructive algorithms, such as block-level ME, by removing faulty MVs, and so on. The main goal of this technique is to effectively use level-wise benefits while learning from selected information to get beyond inherent restrictions. A method based on occlusion reasoning was put out by Won Hee Lee *et al.* [44], who predicted four interpolated frames utilizing the accuracy of estimated motion vector fields generated by a sophisticated optical flow framework [38].

Zhao *et al.* [45] offered an edge-based enhancement of estimated MVFs using hole filling in the MCI unit and edge information from the variable block ME module. Comparing the edge-based component to the traditional optical flow-guided and MSEA methods, the computation overhead of the edge-based component is countered by the visibly high quality of the findings. To reduce accuracy degradation before BME, Li *et al.* [46] created a low-complex version with an advanced EPF that subsamples high-frequency parts of video frames. By assisting BME in reducing mismatched blocks and neutralizing the negative impacts of homogenous forms in texture sections of video frames, the real-time EPF implements edge preservation of fundamental objects.

2.1.3 Kernel-based approach

The kernel-based approaches [32], [31], [26] integrate the image by working across local patches across each pixel rather than using solely pixel-wise information to maintain the local textual characteristics of the frame.

Niklaus *et al.* [19] create pixel interpolation as a local convolution over patches in the input images and blends motion estimation and pixel synthesis into a single step.

To prevent the two-step method from being hampered when optical flow is unreliable because of occlusion, motion blur, and texture deficiency. This method's primary goal is to calculate the ideal convolutional kernel that will be used to synthesize each output pixel in the interpolated images. This approach’s convolution kernel coefficients must be non-negative and sum to one, which is a crucial requirement.

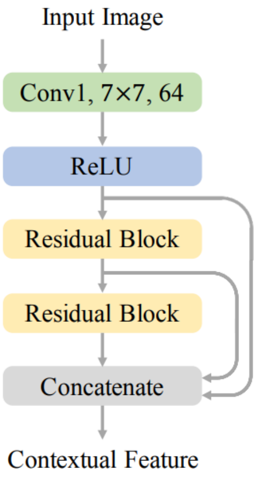


Figure 14: Context extraction network

Bao *et al.* [20] developed a depth-aware flow projection layer to get intermediate flows. The input frames, depth maps, and contextual data were then warped within the adaptive warping layer. Flow estimation, depth estimation, context extraction, kernel estimation, and frame synthesis networks are among the submodules that make up this model. This network produces the output frame via residual learning. This composition of the context extraction system is illustrated in Figure 14 and Figure 15.

By using self-developed frame-warping techniques, many existing methods frequently locate regions with relevant information to correctly assess each output pixel. However, the bulk of present methods have a limited degree of freedom (DoF) and cannot go through the real-time challenges of complex motions. The most recent warping module, dubbed adaptive collaboration of flows (AdaCof), was created by Lee *et al.* [47] to address this problem. It is established on an operation that utilizes any number of pixels in any point. In contrast to SepConv [32], this approach creates the output frame by processing discrete offset vectors and kernel weights for every target pixel.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

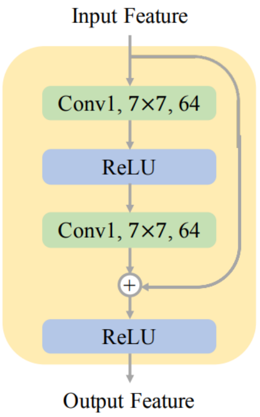


Figure 15: Residual block

In contrast to traditional optical flow approaches [38], [43], [48], it offers a more generalized warping framework and redefines the majority of those as special cases of it. To reasonably construct real-time intermediate frames, the network architecture integrates dual-frame adversarial loss with a fully convolutional neural network advancing over DSepConv [49]. Cheng *et al.* (2020) proposed utilizing additional pertinent pixels to approximate kernels adaptively under deformable separable convolution [49], using a smaller kernel size with pertinent features to handle significant motion. DSepConv uses the encoder-decoder network to extract features. These attributes are utilized for each pixel in the frame to estimate separable kernels, masks, and offsets. In EDSC [50], the developers of DSepConv enhanced their earlier model. They could train with fewer parameters while still getting the same outcomes. They were also the first kernel-based method to successfully produce numerous interpolated frames between two consecutive frames. However, the outcomes for arbitrary time interpolation were inferior to cutting-edge flow-based methods.

2.1.4 Deformable-convolution-based approach

Flow-based and kernel-based approaches shown in Figure 16, combined in the deformable-convolution-based approach [47], [51], [50], [52]. The deformable convolution approach illustrates the benefit of variable spatial sampling, which is utilized in this method. This approach addresses the issues of intricate image texturing and quick scene movement. Low-cost deep learning techniques are very popular that efficiently exploit the color and motion information of high-quality sequences of video changes to interpolation processing techniques, great progress has been made.

Deep learning methods have shown encouraging outcomes on a variety of restricted datasets. Despite this progress, much more work needs to be done to provide outcomes that meet real-time requirements. Creating an effective, intelligent interpolation system is extremely laborious. Numerous difficulties arise from such an idea.

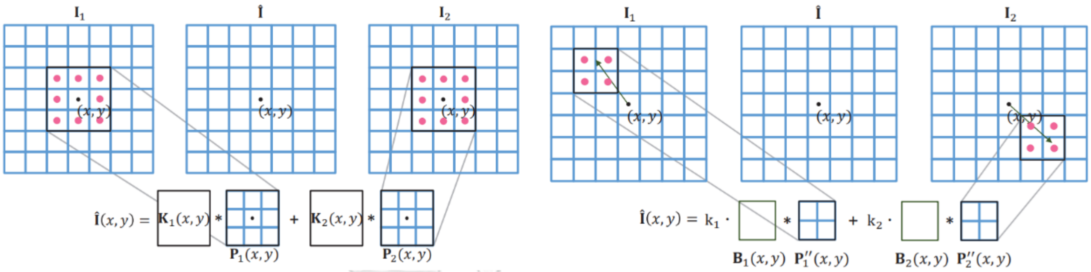

I_1
$\hat{I}$
I_2
(x, y)
$\hat{I}(x, y) = K_1(x, y) * P_1(x, y) + K_2(x, y) * P_2(x, y)$
$\hat{I}(x, y) = k_1 \cdot B_1(x, y) * P_1''(x, y) + k_2 \cdot B_2(x, y) * P_2''(x, y)$

Figure 16: Deformable convolution approach

2.1.5 Visual artifacts and occlusion

These between-frame differences, which are calculated using depth maps [26], [27] may result in viewpoint fluctuation that obscures the details of repeated arrivals of the same activities. To reduce occlusion, several frequently employed datasets involve subjects to perform actions in a constrained and visible background knowledge, resulting in less obstructed but constrained view data gathering. However, interactions in real-world scenarios are inevitably occluded, making it difficult to separate entities in overlapping regions and extract the features of individual objects. As a result, many existing approaches are ineffective.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.2 Video Super-Resolution (VSR)

Several different VSR techniques have been put forth recently. Traditional approaches and deep learning make up most of them. As to Schultz and Stevenson, some conventional techniques only use affine models to estimate the motions (1996). For VSR, Protter *et al.* [53] and Takeda *et al.* [54] utilize non-local mean and 3D steering kernel regression, respectively. Liu and Sun [55] recommended a Bayesian method to reconstruct high-resolution frames to simultaneously estimate the underlying motion, blur kernel, and noise level. Ma *et al.* [56] utilize the probability maximization (EM) methodology to approximate the blur kernel and direct the reconstruction of HR frames. These explicit HR video versions are still inadequate to accommodate different video consequences.

SR methods established on deep learning are the topic of substantial research due to deep learning's outstanding performance in several fields [57].

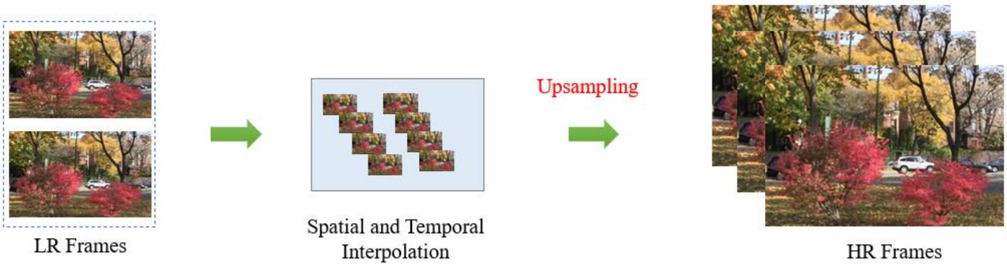


Figure 17: Basic concept of VSR tasks

There have been many deep neural network-based video superresolution techniques developed. A taxonomy for VSR methods is illustrated in Figure 18.

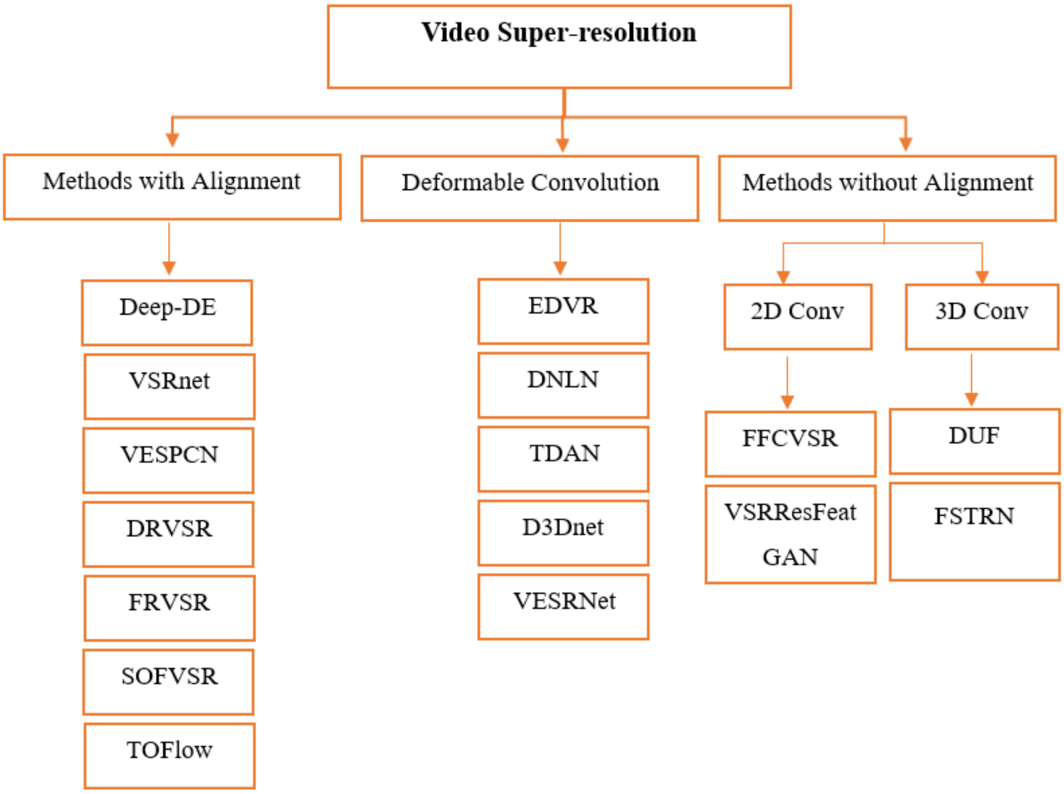


Figure 18: A taxonomy for VSR methods

2.2.1 Methods with Alignment

Most methods for VSR alignment use motion estimation and compensation procedures. Motion estimation is particularly managed to obtain motion information within the frame. In contrast, motion compensation is employed to warp the frames after the motion information from the inter-frames and align them. The optical flow method is used to carry out most motion estimating procedures.

a. Deep-DE

Figure 19 describes the two stages of the deep draft-ensemble learning methodology Deep-DE [58]. Four convolutional layers make up the CNN in Deep-DE: the first 3 stages are of deconvolution layers, while the fourth one is a typical convolution layer. The kernel dimensions of these layers are 11 by 11, 1 by 1, 3 by 3, and 25 by 25, and there are correspondingly 256, 512, 1, and 1 channels.

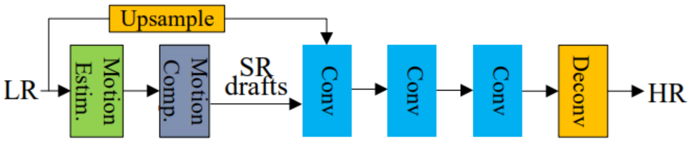


Figure 19: The architecture of Deep-DE

b. VSRnet

The network design of VSRnet [59], based on the image SR method, is illustrated in Figure 20. Three convolutional layers and three motion estimation and compensation modules make up most of VSRnet. Each convolutional layer, except the final one, which is followed by the rectified linear unit (ReLU).

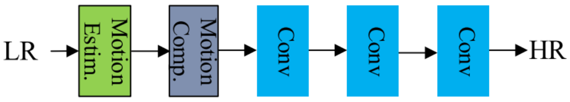


Figure 20: VSRnet architecture

The quantity of input frames is the primary distinction between VSRnet and SRCNN. In other words, while VSRnet employs a series of subsequent, adjusted frames, SRCNN only accepts a single input frame.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

c. VESPCN

A spatial motion compensation transformer (MCT) segment is recommended by the video-efficient sub-pixel convolutional network (VESPCN) [60] for motion estimation and motion compensation. In Figure 21, the modified frames are input into various convolutional layers for extraction of features and fusion. At end, a sub-pixel convolutional layer for upsampling is utilized to generate the SR results.

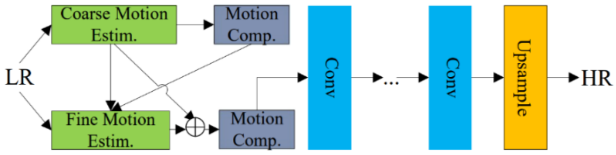
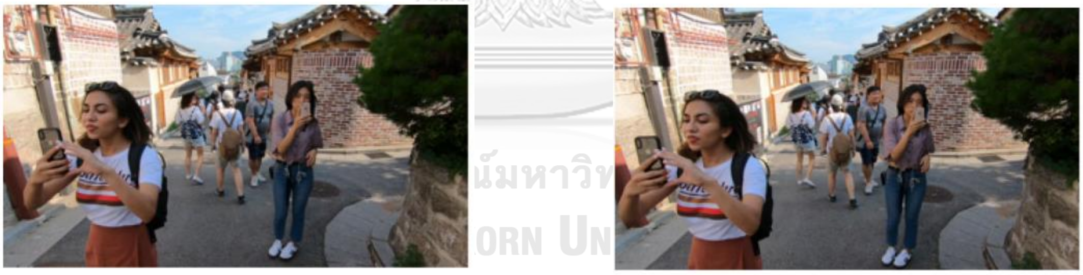


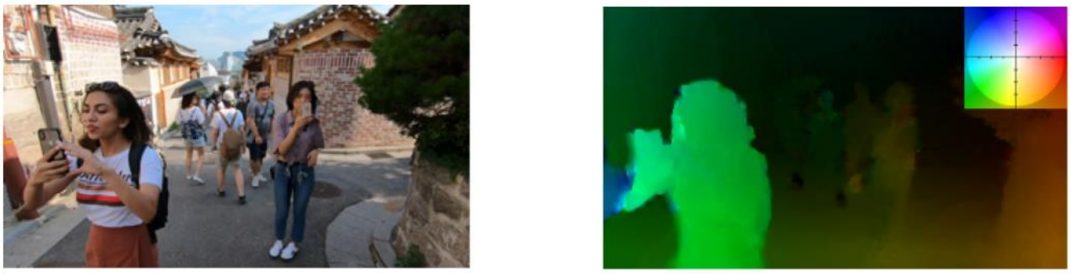
Figure 21: The network architecture of VESPCN

d. DRVSR

Corresponding to the optical flow information, the detail-revealing deep video super-resolution (DRVSR) [61] illustrated in Figure 23, methodology suggests a sub-pixel motion compensation layer (SPMC) that can carry out the upsampling and motion compensation procedures at the same time for neighboring input frames.



a. A targeted frame b. Its neighboring



c. Compensated image d. Estimated optical flow image

Figure 22: Motion estimation and compensation

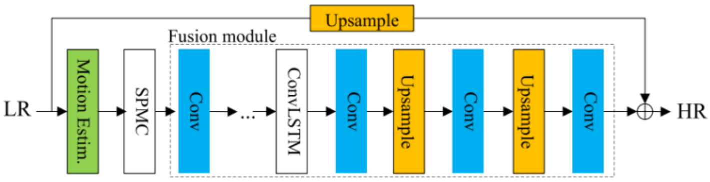


Figure 23: Architecture of DRVSR

e. FRVSR

Frame recurrent video super-resolution (FRVSR) [62] mainly recommends to utilize the formerly inferred HR approximate to super-resolve the following frame to get temporally consistent outcomes. An optical estimate network is used in the detailed implementation to compute the optical flow from the preceding frame to the target frame. The LR flow is upsampled utilizing bilinear interpolation to the similar resolution as the HR video.

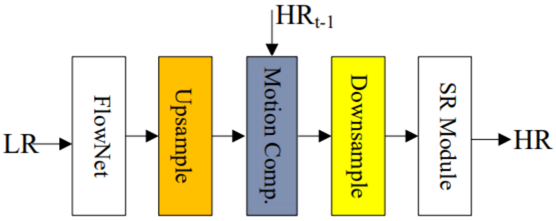


Figure 24: Architecture of FRVSR

f. SOFVSR

It is recommended to super-resolve LR expected optical flow for video super-resolution (SOFVSR) to get exceptional SR performance. The optical flow reconstruction network (OFRnet) is utilized to estimate the optical flow among frames ultimately generates HR optical flow. After that, a space-to-depth transformation is used to change the HR optical flow into the LR optical flow. The LR optical flow warps the adjacent frames to create the target frame lines up with its neighbors.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

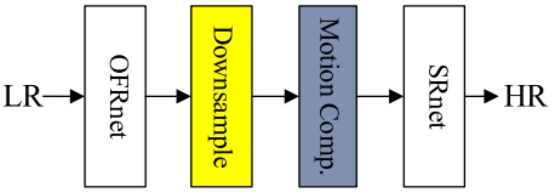


Figure 25: Architecture of SOFVSR

g. TOFlow

The task-oriented flow's (TOFlow) [24] architecture is shown in Figure 26. TOFlow operates SpyNet as the network for estimation, and a spatial transformer network (STN) is employed to warp the neighboring frame to calculate optical flow. The image processing segment for the VSR task comprises of 4 layers of convolution, having kernel sizes of 9 by 9, 9 by 9, 1 by 1, and 1 by 1, and channel counts of 64, 64, 64, and 3, respectively.

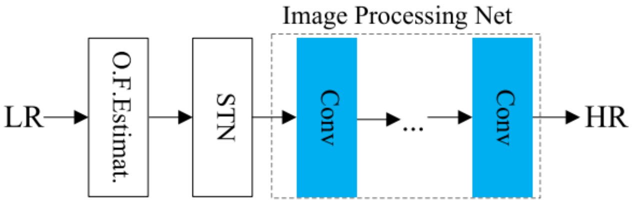


Figure 26: Architecture of TOFlow

2.2.2 Methods with Deformable convolution

Dai [63] originally presented the deformable convolutional network, and the enhanced version was proposed in 2019. Because it is customary in conventional CNNs to utilize fixed geometric formations in every layer, the network's capability to model geometric transformations is restricted. Figure 27 illustrates the deformable convolution for feature alignment. By projecting the target feature maps onto the nearby feature maps, further convolutional layers can be used to achieve offsets.

จุฬาลงกรณ์มหาวิทยาลัย

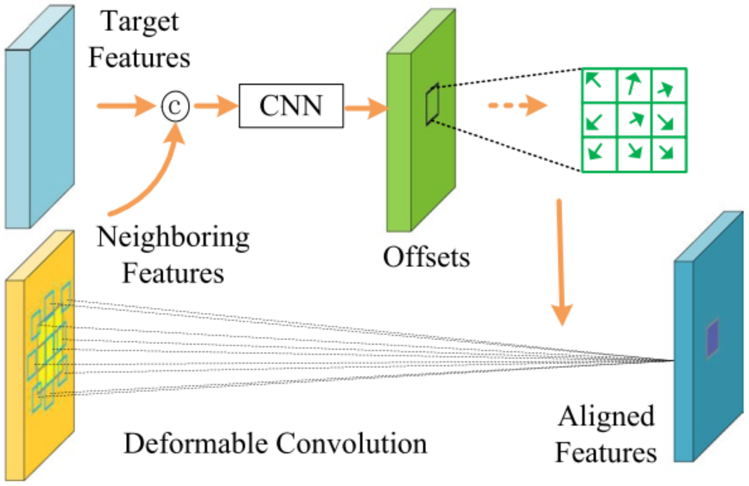


Figure 27: Deformable convolution

a. EDVR

The winning model in the NTIRE19 Challenge is the improved deformable video restoration (EDVR) [64], which is shown in Figure 28. The temporal-spatial attention (TSA) fusion module and the pyramid, cascading, and deformable (PCD) alignment segment is utilized by EDVR to proficiently fuse several frames and solve large motions in videos, respectively.

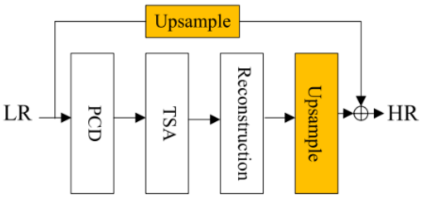


Figure 28: The network architecture of EDVR

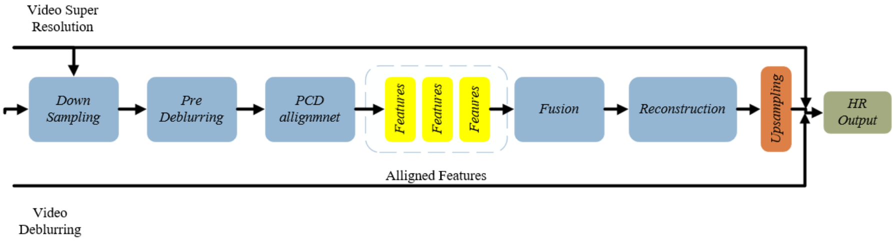


Figure 29: Overview of Enhanced Deformable Video Restoration (EDVR)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

b. DNLN

Consisting on deformable convolution [63], [65] and non-local networks, the deformable non-local network (DNLN) [66] creates an alignment segment and a non-local attention segment [67], respectively. The alignment segment uses the original deformable convolution's hierarchical feature fusion module (HFFB) [68] to produce convolutional parameters. Additionally, DNLN employs numerous deformable convolutions in a cascaded approach, enhancing inter-frame alignment.

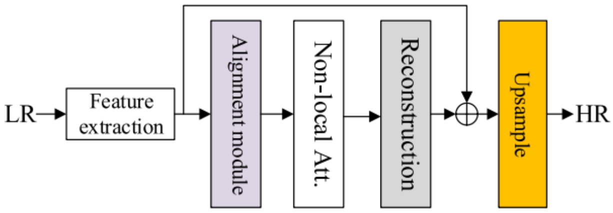


Figure 30: Architecture of DNLN

c. TDAN

The targeted and neighboring frames are subjected to deformable convolution by the temporally deformable alignment network (TDAN) [16], which accomplishes matching offsets. The adjacent frame is then offset-warped to line up with the target frame. A feature extraction section, a deformable convolution section, and a reconstruction section make up the three sections of TDAN.

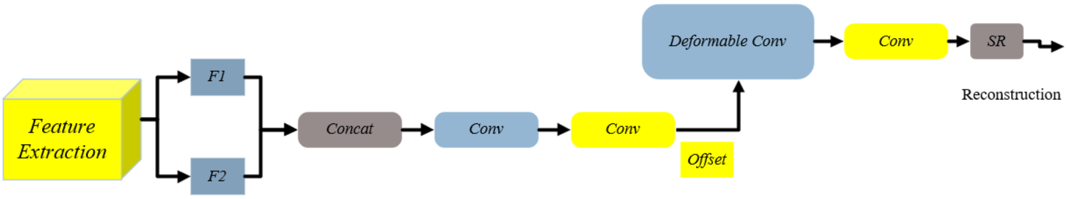


Figure 31: Overview of Temporally Deformable Alignment Network (TDAN)

d. D3Dnet

Figure 32. illustrates the layout of the deformable 3D convolution network (D3Dnet) [69]. To achieve strong spatiotemporal feature modeling capabilities, D3Dnet suggests 3D deformable convolution. The inputs are fed into a 3D convolutional layer to generate features, and then to Residual Deformable 3D Convolution (ResD3D) blocks to compensate for motion information and capture spatial information.

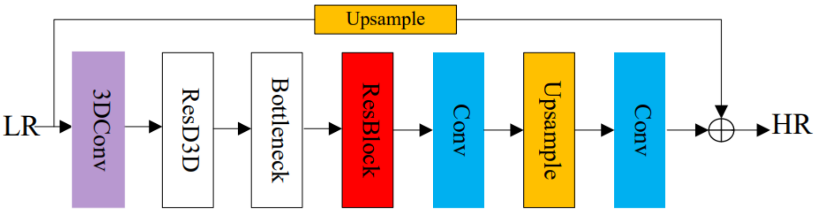


Figure 32: Architecture of D3Dnet

e. VESR-Net

Video enhancement and super-resolution network (VESR-Net) [70], illustrated in Figure 33. A feature encoder, a fusion section, and a reconstruction section makes most of the VESR-Net. Separate NL can fuse the data through frames of video and pixels in each frame with a small number of parameters and a lighter network than the standard non-local architecture [67]. The reconstruction module employs CARBs followed by a feature decoder for upsampling, while the upsampled module is implemented via a sub-pixel convolutional layer. Additionally, it generates the super-resolved frame by blending it with the LR target frame by applying bicubic interpolation.

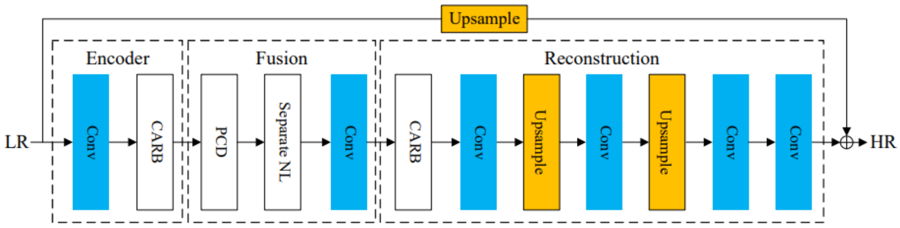


Figure 33: VESRNet architecture

2.2.3 Methods without alignment

These methods which are working without alignment cannot align neighboring frames for VSR. For feature extraction, these algorithms primarily use spatial or spatiotemporal information.

a) 2D convolution methods

The frames are directly fed to the 2D convolutional network to perform processes such as feature extraction, fusion, and SR to perform alignment tasks. The network is forced to learn the correlation information inside frames, this might be a straightforward solution to the VSR problem. The exemplary techniques are FFCVSR [71] and VSRResFeatGAN [72].

จุฬาลงกรณ์มหาวิทยาลัย

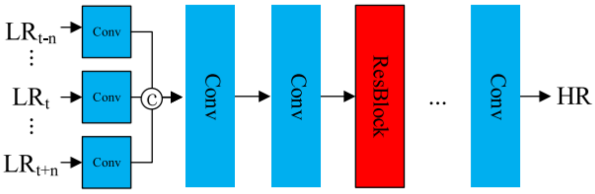


Figure 34: The architecture of the generator in VSRResFeatGAN

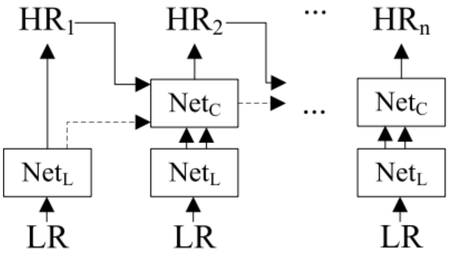


Figure 35: The architecture of FFCVSR

b) 3D convolution methods

The 3D convolutional module [73] acts on the spatiotemporal domain. As a result, the correlations between frames are considered while processing video sequences, which is advantageous. The representative 3D convolution techniques for VSR include DUF [74].

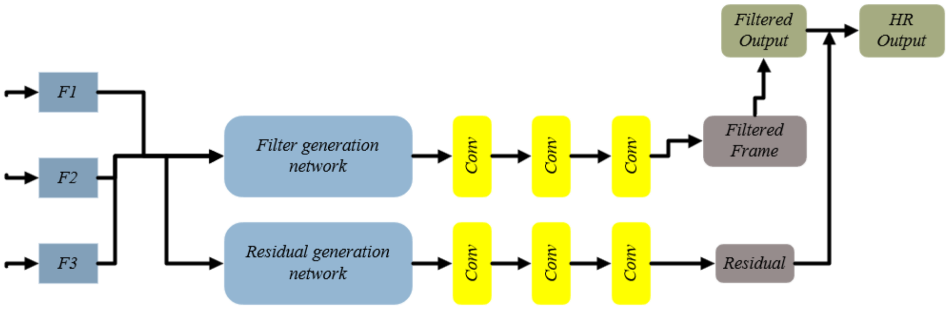


Figure 36: Overview of Dynamic Up-sampling Filters (DUF)

2.2.4 Recurrent back-projection network

A Recurrent Back-projection Network (RBPN) that acquires the flow maps of multi-frames and concatenates them with LR video frames was proposed by Haris et *al*. [75]. However, it is difficult to establish precise flow in RBPN and unpleasant artifacts appear in aligned frames. Figure 37 illustrates the overview of RBPN.

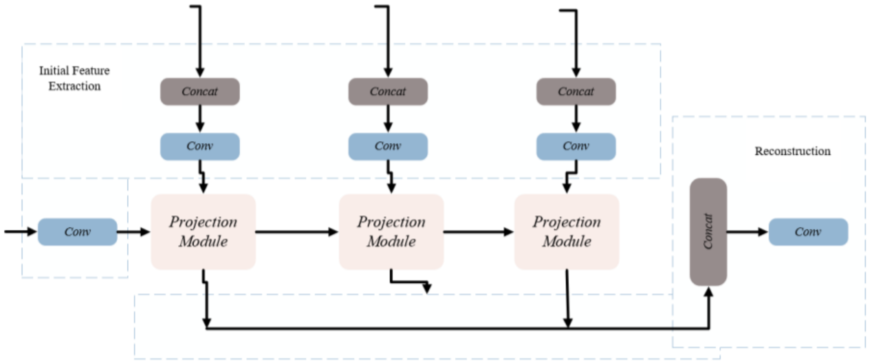


Figure 37: Overview of RBPN

2.2.5 Convolutional LSTMs

Convolutional LSTMs [76] (ConvLSTM) are applied to VSR algorithms because of the latest CNNs' ability to simplify the learning of the sequence-to-sequence (S2S) model and subsequently improve utilizing the temporal information. Although using ConvLSTM considerably enhances the VSR outcomes, the Recurrent Neural Network

(RNN) performs poorly in big and complex motions of the frames due to the absence of explicit temporal alignment. Xiang *et al*. [19] demonstrate a novel ConvLSTM approach by embedding it with an explicit state update cell to improve the efficiency of VSR. This means that the one-stage space-time VSR concurrently learns the spatial SR and the temporal feature interpolation without needing the supervision of intermediate LR frames.

2.3 Channel Attention (CA)

In convolutional neural networks, we create a channel attention map by leveraging the link between features across channels. A feature map's channels are feature detectors, so channel attention focuses on "what" is significant input image. We reduce the input feature map's spatial dimension to compute the channel attention effectively.

It is widely acknowledged that human vision depends heavily on attention [77], [78]. Human vision does not seek to process an entire scene at once, an important characteristic of the human visual system. Instead people use a series of fragmentary glimpses to better understand the visual organization and focus on relevant portions [79]. Recently, various attempts [80], [81] have been made to add attention processing to enhance CNN performance in challenging classification tasks.

2.3.1 Channel attention module

A feature map's channel is regarded as a feature detector [82]. When provided an input image, channel attention concentrates on "what" is significant. We reduce the input feature map's spatial dimension to compute the channel attention effectively. Average pooling has been widely used to aggregate spatial data so far.

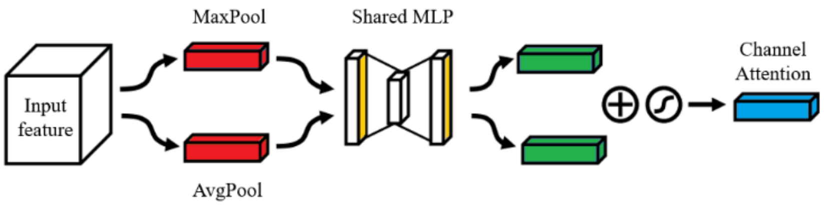


Figure 38: Channel Attention Module

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

The key-value set pair at the side of the output and the mapping query is the defined functions of attention, wherever the values, keys, query, and output are the vector quantities. The output is determined by the addition of values having some weight. The weight of every value is determined by the compatibility function of a query with its key corresponding to it.

2.3.2 Scaled Dot-Product Attention

This attention is illustrated in Figure 39. The queries, dimension values d_v, and the dimension keys d_k are attention inputs. The dot product is determined by all the keys with the query and divided by each dimension key after that. Each carries the identical input sequence that has been enhanced and encoded with positional information in the encoder stage. The queries and keys represent the identical target sequence. Values sent into the first attention block on the decoder side following this, which would also have been enhanced and embedded with positional information. The decoder's second attention block gets the encoder output as keys and values and the first attention block's normalized output as queries. The dimension of the keys and queries is represented by d_k, whereas the dimension of the values is represented by d_v. These queries, keys, and values are sent as inputs to the scaled dot-product attention, which then computes the dot-product of the queries with the keys. The attention scores are then created by scaling the result by the square root. After feeding them into a SoftMax function, a collection of attention weights is obtained. The values are finally scaled using the attention weights via a weighted multiplication process. The whole procedure can be mathematically stated Q, K, and V are the queries, keys, and values, respectively. In the end, the SoftMax function will be applied to attain the weight of the values [83].

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

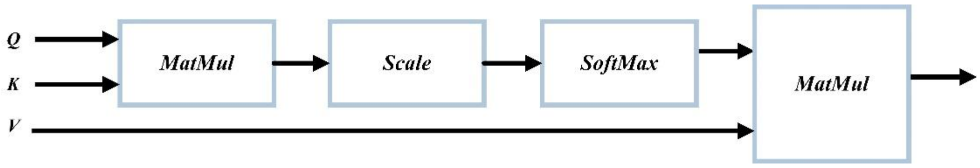


Figure 39: Attention with Scaled Dot-Product

$$Attention\ (Q, K, V) = softmax * \left[\frac{QK^T}{\sqrt{d_k}}\right] * V \qquad (2.1)$$

Generally, the attention function is determined on the sets of queries, at the same time, are wrapped into a matrix $\boldsymbol{Q}$. Similarly, the values and the keys are also wrapped in the shape of matrix $\boldsymbol{K}$ and $\boldsymbol{V}$.

2.3.3 Multi-Head Attention

This attention is comprised of various single-attention functions. It also performs attention to the queries, values, and keys. It is beneficial after linearly projecting the keys, queries and values, n times with several linearly projected dimensions, i.e., d_k, d_q and d_v respectively. The attention function is applied to each projected type of key, query, and value parallelly, yielding the output values of dv dimension. In the end, the result is a computer after applying concatenation [83].

The input was duplicated as vectors for the queries, keys, and values. These Query, Key, and Value inputs were each passed through a fully linked layer to decrease the parameters. A correlation matrix was generated using the condensed Query and Key inputs; it was then scaled down and placed through SoftMax to produce a weighted correlation matrix. Contextualized embedding vector that can be supplied to the decoder was created by multiplying the reduced value vector by this correlation matrix. This contextualized vector representation can be improved further.

Additionally, we can divide our single input vector into numerous smaller chunks. Let's say divide the embedding vectors of 768 sizes into 4 blocks. These learned embeddings are what we employ in the embedding layer. Each word has a relevant meaning or concept attached to it. Different concepts are present in various embedding positions. After the input is divided into blocks, each block can present a notion. The final contextualized vector blocks will have more precise control over the notion they represent with the other words in the input when the attention model is applied on top of this. The Multi-head Attention Model refers to this. The input is divided into several heads, and the attention model runs on each of these heads separately. Information collection from various subspaces with different positions is allowed to the model by the multi-head attention. Expanding the single attention head is expressed below in the equation and illustrated in Figure 40.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

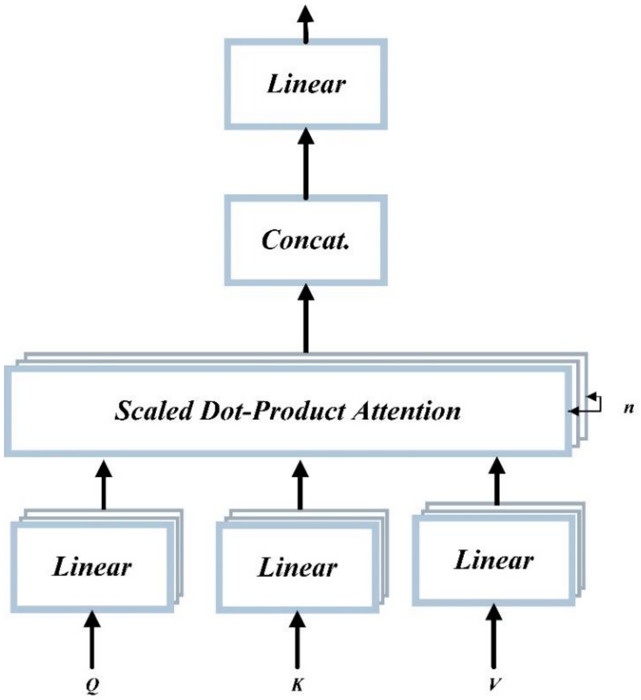


Figure 40: Attention with Multi-Head

$$Multi - head\ (Q, K, V) = conc.(head_1, \ldots \ldots, head_n) * W^o, \tag{2.2}$$

Where.,

$$head_i = Atten.(QW_i^Q, KW_i^k, VW_i^v) \tag{2.3}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.4 Space-Time Video Super-Resolution (STVSR)

The objective of STVSR is to simultaneously improve the spatial and temporal resolution of video. Video spatial super-resolution (S-SR) and temporal super-resolution (T-SR) have performed remarkably well thanks to the development of CNN. To execute STVSR, it is thus simple to do T-SR and S-SR in that order (two-stage). However, spatial augmentation and temporal interpolation are inextricably linked and may work best together. It is challenging to benefit from this attribute when it is being processed separately. The two-stage method also frequently contains many parameters and is difficult to implement. STVSR aims to super-resolve the LR frames into HR frames while considering spatial fusion and temporal alignment. As a result, it is crucial to properly utilize the temporal relationships between various frames. Shechtman *et al.* [84] groundbreaking space-time SR tackled the issue of simultaneous spatial and temporal super-resolution.

Mudenagudi *et al.* [85] developed graph-cuts [86] optimization and maximum a posteriori-Markov Random Field [87] to address the STVSR model's reconstruction problem. Although STVSR models have improved in the ways mentioned above, they still have high computational costs and cannot simulate complex space-time visual patterns.

Recent learning-based STVSR models combine the spatial and temporal challenges into a single-stage framework [19], [20], [21], [88] to overcome these problems. Based on the single-stage framework, the U-net design is used in the Kim *et al*. [88] model to offer a multi-scale spatial-temporal loss. To improve the frame interpolation process, Haris *et al.* [20] single-stage model is applied with the pretrained optical flow. Deformable convolution and bidirectional deformable ConvLSTM were employed by Xiang *et al.* [19], and a unified STVSR model was proposed to improve interpolation between intermediate frames and boost global temporal correlations.

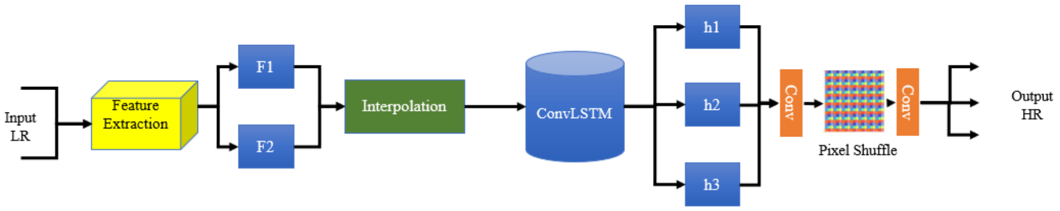


Figure 41: Overview Zooming Slow-Mo

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

Xu *et al.* [21] presented a locally temporal feature comparison module and improved performance to address the local motion feature limitation of the Xiang *et al.* [19] model. The difficulty of extracting local motion cues, however, still exists.

CHAPTER 3

PROPOSED METHOD

The fundamental structure of the suggested methodology is explained in this section. Following that, we arrive at an empirical conclusion on the multi-head channel attention (MHA) block by analyzing the impact of careful feature temporal interpolation of LR frames.

3.1 Overall Framework

We first acquire the following visual features: F_{t-1}^{L} and F_{t+1}^{L} by feature extraction, as illustrated in Figure 42, from I_{t-1}^{L} and I_{t+1}^{L} respectively. The deep consecutive channel attention module receives the retrieved LR features as input, and multi-head attention is applied. CNNs make considerable use of the channel attention strategy. Assume that $X \in \mathbb{R}^{C\times H\times W}$, which is an image feature in a network. Channel count as a whole is represented by C. The dimensions of a feature are H and W, respectively. The multi-head attention mechanism takes the LR feature frames and extracts the important data. The module for attentive feature temporal interpolation synthesizes the feature map F_{t}^{L}, corresponding to the missing intermediate feature map. In addition, to better exploit temporal information, a deformable ConvLSTM processes the temporally consecutive feature maps $\{F_{t}^{L}\}_{t=1}^{2n+1}$. The quality of the features is currently high enough to use convolution. After applying the PixelShuffle [60] to features, high-resolution frames are rebuilt.

CHULALONGKORN UNIVERSITY

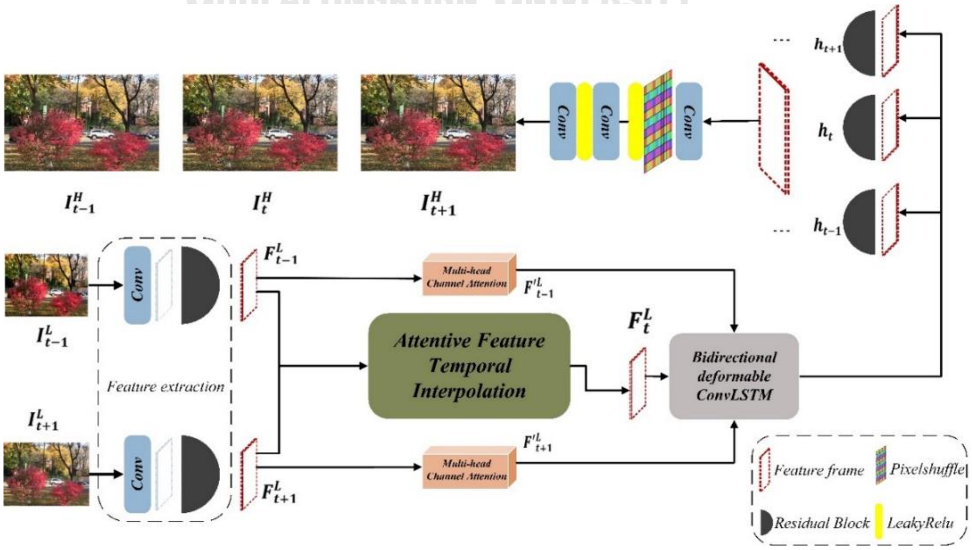


Figure 42: The framework of Deep Consecutive Attention Network for VSR (DCAN)

3.2 Deep consecutive attention model

Two steps make up the deep consecutive attention model: (i) the early stage of feature extraction and (ii) the attentive feature temporal interpolation stage. Below is a summary of these phases.

3.2.1 The early stage of feature extraction

From I_{t-1}^{L} and I_{t+1}^{L}, respectively, which include LR frames, this segment extracts visual features F_{t-1}^{L} and F_{t+1}^{L}. The 3-D feature maps having spatial information of frames are extracted with the help of feature extraction module. The network for feature extraction consists of a layer of convolutional and residual blocks, designated k_1 . After this 1 × 1 convolutional layer performs linear adaptation to make the desired dimension *d* of the features. Given $X \in (0, 255)^{3\times H\times W}$, which is the input feature maps having the size of visual feature as $\boldsymbol{V} \in \mathbb{R}^{w\times h\times d}$, the *(w, h)* represents the W/32 and H/32. Then, these feature maps flatten into 2D as $V_f = (v_1, v_2, \ldots., v_l)$. The $l = w \times h$ and the $(v_{1, 2, \ldots, l}) \in \mathbb{R}^{d}$contains the spatial information.

By paying attention to the feature maps in this design, multi-head attention supports the task to get attentive features providing the meaningful information of the feature maps. Moreover, these maps contain supplementary data and capture the subtleties of elements absent from supporting frames. As a result, feature temporal resolution aids in the effective extraction of characteristics of the intermediate missing frame.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.2.2 Multi-head Attention

In a multi-head attention module, the attention mechanism operated simultaneously in a parallel manner for all input sequences. The outputs of the attention mechanisms concatenate together and transform linearly into the desired dimension. Generally, the multi-head concept helps our model to have attention to the LR features in different shorter and longer feature aspects.

We used scaled dot-product attention to concentrate on the important spatial regions for frame representation. The advantage of this architecture is that the representation vectors of a frame maintain the important information throughout the scene. The multi-head attention module is applied over the feature extractor of each modality. Our objective is to learn a series of attention weight vectors, each of which focuses on a distinct subset of spatial features and is used to refine significant spatial characteristics.

Step No 1:

We use three different linear layers to create Query (Q), Key (K), and Value (V) and these linear layers have their own weights. The matrix of shape of feature is generated multiplying the embedding size of the features with the weight matrices. Then, this input is fed to the linear layers for the creation of Q, K, and V matrices. The information is spread across each head to apply attention across every single head. Basically, the Q, K, and V are logically divided into different matrices not physically because they contain the same information. This is done by using a single data matrix.

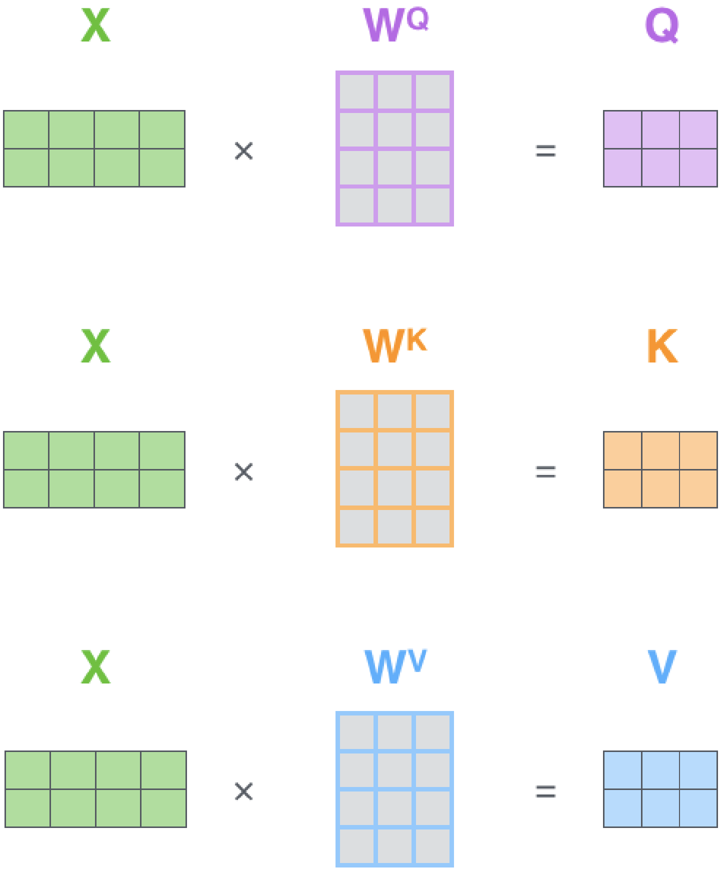


Figure 43: Query, Key and Value

Step 2:

At this stage, we have our input vectors for the attention mechanism to calculate the score of the attention. This mechanism is illustrated in figure 44.

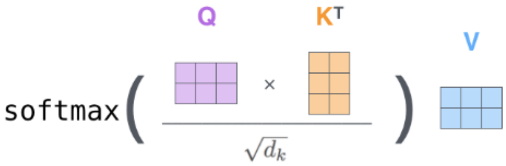


Figure 44: Calculation of attention Score

The attention score is calculated by the dot product between the *Q* and the transpose of *K* which is $(Q * K^{T})$.

Step 3:

We use the $\sqrt{d_k}$ to have stable value of attention otherwise, the gradient of features will become very small. The attention score will be divided with the dimension of the *K* while taking the square root of it. This mechanism helps us to have more stable gradients. Then we pass it through SoftMax by multiplying the attention score with it to normalize the attention score.

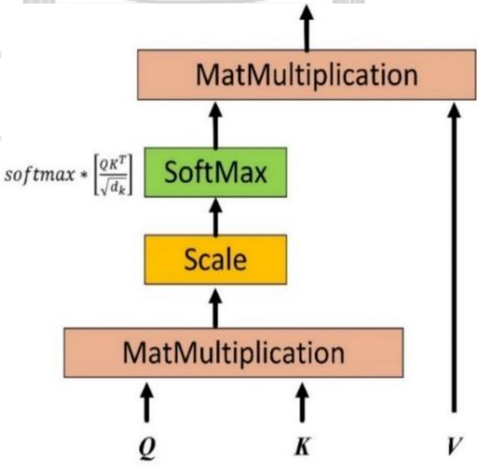


Figure 45: Scaled Dot-Product Attention

Here now we have separate scores of attentions of individual head. This needs to be merged into a one single score. For this first reshape the matrix of attentions core by swapping of the sequence and the head dimensions i.e., (Batch size, Height × Width, Head). Basically queries, keys, and values expressed as head h individually learned linear projections. These head h projected queries, keys, and values are simultaneously fed into attention. The final output is created by concatenating the outputs from attention and learned linear projection. Multi-head attention is described as learning linear transformation using fully connected linear layers.

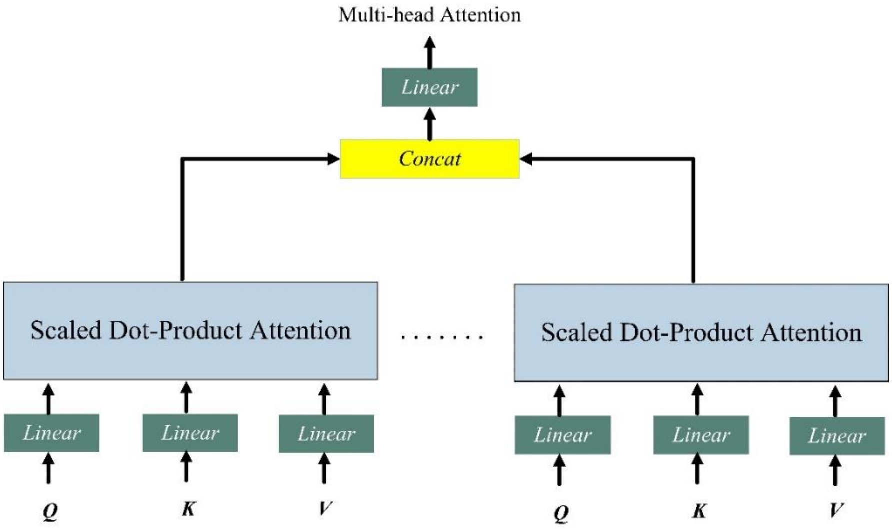


Figure 46: Multi-head attention

Mathematically it will be represented as:

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

- Query $\boldsymbol{q} \in \mathbb{R}^{d_q}$
- Key $\boldsymbol{k} \in \mathbb{R}^{d_k}$
- Value $\boldsymbol{v} \in \mathbb{R}^{d_v}$
- Attention head $\mathbf{A}_i$ (i = 1, . . . , h)

$\boldsymbol{h_i}$ will be computed as:

$$h_i = \left(\boldsymbol{W}_i^{(q)}, \boldsymbol{W}_i^{(k)}, \boldsymbol{W}_i^{(v)}\right) \in \mathbb{R}^{hd_v \times d_{model}} \tag{3.1}$$

The output of multi-head has a linear transformation from the learnable parameters of each head $\boldsymbol{h}$ after concatenation. In summary, three identical inputs are sent to the multi-head attention module. Three trainable matrices (Linear layers) are used to create three vectors for each feature (query, key and value). These vectors fill the matrices Q, K, and V, one characteristic after another. Each query vector is compared to all the keys, and a key represents a vector. A query should be similar to the keys for terms that have some link for affinity, or connection with the query itself. In the QK^T matrix, this similarity is represented by the dot products of the rows and columns. To prevent an excessive increase in the size of the products, a division by scale, the square root of dim_head, is used. We employ parallel layers of attention or heads $h = 8$. For every layer $d_k = dv = d_{model}/ h = 64$. The total computing cost is comparable to that of single-head attention with full dimensionality because of the lower dimension of each head. Using single matrix operation, the computations of all the heads are calculated instead of several operations. By doing this with the help of a few linear layers the model remains simple and provides efficient computations.

Applying multi-head attention on extracted features of F^L_{t-1} and F^L_{t+1} which will become the attentive feature F'^L_{t-1} and F^L_{t+1}.

$$Multihead\ (Q,K,V)\ [F^L_{t-1},\ F^L_{t+1}] \Rightarrow F'^L_{t-1}, F'^L_{t+1} \tag{3.2}$$

3.2.3 Attentive feature temporal interpolation

Using the output feature maps from the LR sequences, F^L_{t-1} and F^L_{t+1}, as input to create the intermediate feature map F^L_t. To generate the F^L_t, the interpolation function *f*(.) for one stage STVSR can be represented as follows in equation (3.3).

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$F^L_t = f(F^L_{t-1}, F^L_{t+1}) \Rightarrow B(S_{t-1}(F^L_{t-1}, \varphi_{t-1}), S_{t+1}(F^L_{t+1}, \varphi_{t+1})) \tag{3.3}$$

The sampled features are aggregated using the blending function B(.). The sampling functions are $S_{t-1}(\cdot)$ & $S_{t+1}(\cdot)$, and the sampling parameters are φ_{t-1} & φ_{t+1}, respectively. This makes it easier to distinguish between characteristics that move forward and backward. In contrast, it is impossible to compute forward and backward motion information using F^L_t. To alleviate this issue, the motion information is carefully interpolated and an interim feature map, F^L_t, is generated. This module can handle high motion rates in videos. The sampling functions have different weight sizes but use the same network design. To understand this $S_i(\cdot)$ taken as an illustration. As input, it operates feature maps with LR frame F^L_{t-1} and F^L_{t+1} , to determine an offset for testing the F^L_t .

$$\Delta p_{t-1} = g_{t-1}([F^L_{t-1}, F^L_{t+1}]) \tag{3.4}$$

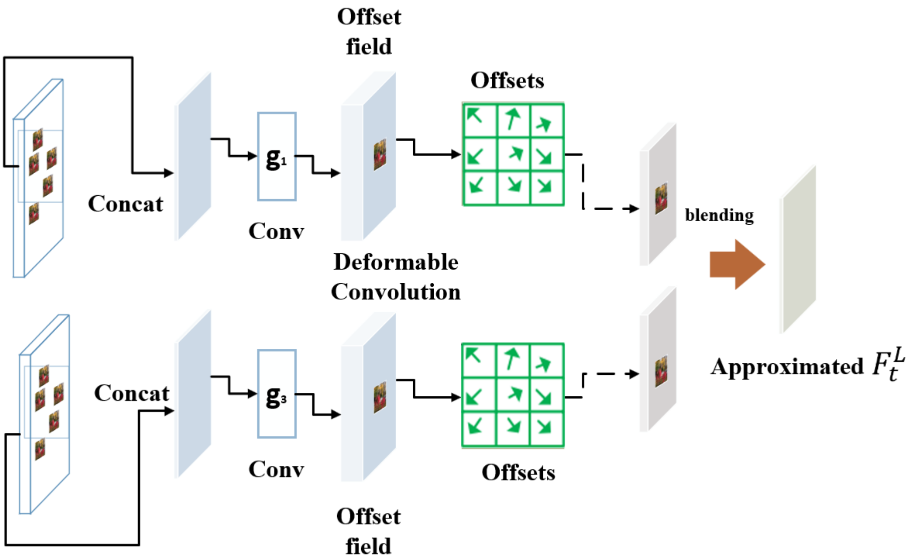


Figure 47: Attentive feature temporal interpolation module

here Δp_{t-1} is learnable offset and g_{t-1} represents a common function of various convolution layers. The following equation expressed the deformable convolution for the learned offset Δp_{t-1}.

$$S_{t-1}(F^L_{t-1}, \varphi_{t-1}) = DConv(F^L_{t-1}, \Delta p_{t-1}) \tag{3.5}$$

Likewise, offset for $\Delta p_{t+1} = g_{t+1}([F^L_{t+1}, F^L_{t+1}])$ with φ_{t+1} as sampling parameter. The following equation expressed the deformable convolution for the learned offset Δp_{t+1}.

$$S_{t+1}(F^L_{t+1}, \varphi_{t+1}) = DConv(F^L_{t+1}, \Delta p_{t+1}) \tag{3.6}$$

Simple linear B(.) is used to blend the sampled features:

$$F^L_t = \alpha * S_{t-1}(F^L_{t-1}, \varphi_{t-1}) + \beta * S_{t+1}(F^L_{t+1}, \varphi_{t+1}) \tag{3.7}$$

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

α , β = 1 × 1 learnable kernels of convolution and $*$ = operator of convolution.

Intermediate feature maps $\{F_{2t}^{L}\}_{t=1}^{n}$ can be obtained after applying the function of deformable temporal interpolation to $\{F_{2t-1}^{L}\}_{t=1}^{n+1}$

3.3 Bidirectional deformable ConvLSTM

To generate an HR slow motion video frame with sequence-to-sequence mapping, we got the successive attentive feature maps: $\{F_{t}^{L}\}_{t=1}^{2n+1}$. Previously, in different restoration tasks [89], [61], [64], it is confirmed that temporal information is very important. Before reconstructing HR frames from adjacent frames, we accumulate temporal contexts. ConvLSTM [76] is a standard 2-Dimensional sequence data modeling procedure, and it is adapted to execute temporal accumulation. Given that at the time (t), the cell state c_t and the hidden state h_t, updated by the ConvLSTM with:

$$h_t , c_t \ = \ ConvLSTM(h_{t-1}, c_{t-1}, F_t^L) \tag{3.8}$$

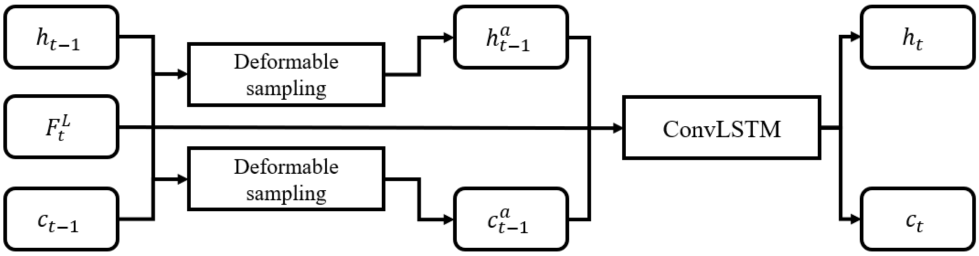


Figure 48: Deformable ConvLSTM

After an in-depth study, it is found that the ConvLSTM has the inadequate aptitude to handle large motions in video sequences. Because it only tacitly captures motion among the previous states: h_{t-1} and c_{t-1}. This creates unembellished temporal divergence among h_{t-1}, $c_{t-1}$and F_t^L. As a result, instead of producing meaningful global temporal contexts it promulgates incompatible "noisy" content.

Therefore, the reconstructed HR frame will experience severe annoying artifacts. Deformable alignment is set as a state cell in the ConvLSTM to eliminate large motion video sequence problems and get efficient global temporal contexts.

$$h_t, c_t = ConvLSTM(h_{t-1}^a, c_{t-1}^a, F_t^L) \tag{3.9}$$

Furthermore, to completely delve into temporal information, bidirectional deformable ConvLSTM is adapted [90]. Feature maps that are temporally inverted are encouraged to use in the deformable ConvLSTM by concatenating the hidden states from backward and forward pass to get the result in the shape of the final hidden state ${h_t}^2$ for HR frame reconstruction. Moreover, bidirectional deformable ConvLSTM is used to improve temporal information, to deal with the successive feature maps: $\{F_t^L\}_{t=1}^{2n+1}$.

3.4 Frame Reconstruction

We obtained successive feature maps to construct HR video slow-motion frames with sequence-to-sequence mapping. A temporally synthesized shared network takes each hidden state as the input and outputs the corresponding high-resolution frame. Bidirectional deformable ConvLSTM enhances temporal information. The residual block k_2 has been stacked by the network to help it understand deep features. Reconstruction uses the upscaling module PixelShuffle [60]. The time and space super-resolution problems in STVSR are intra-related. This method can be trained and comprehend spatial-temporal interpolation at the same time.

3.4.1 Pixel Shuffle

The traditional methods result in new pixel information being formed when an image is enlarged (spatially, along the width and height), which frequently worsens the image quality and produces a blurred image. One of the newest layer types in contemporary deep-learning neural networks is the pixel shuffle layer. Its use is closely related to single-image super-resolution (SISR) research, which studies a group of techniques designed to create a high-resolution image from a single low-

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

resolution one. The first SISR neural networks begin with a bicubic up-sample pre-processing of the input low-resolution image. The model is then fed an image with the same dimensions as the desired output to improve resolution and fix details. In this way, although the quantity of image processing required is less, the number of parameters and additional computational power required by the training section increase (by a factor equal to the square of the desired up-sample scale). To overcome this difficulty a Pixel Shuffle conversion is also identified as *sub-pixel convolution.*

In our framework, the pixel shuffle is used to enlarge the features reconstructed with the scale factor 4. The pixel shuffle shows efficient and superior performances due to rearranging the feature map without losing information and the padding effect. The sub-pixel convolution combines every single pixel on a multi-channel feature into an independent pixel on an image. In the picture below, you can see an illustration in Figure 48 of this transformation:

Let’s assume that the feature map with the dimensions H and W is up-sampling on a scale of t-4. This model groups the feature map into sets of $t^2 = 4$ channels C. Then rearrange each group into a 4 × 4 block of the pixels. Finally, the output size is ($H \times r$, $W \times r$). In other words, the tensor shape ($C \times t^2, H, W$) is rearranged to the tensor shape of ($C, H \times r, W \times r$) without losing information.

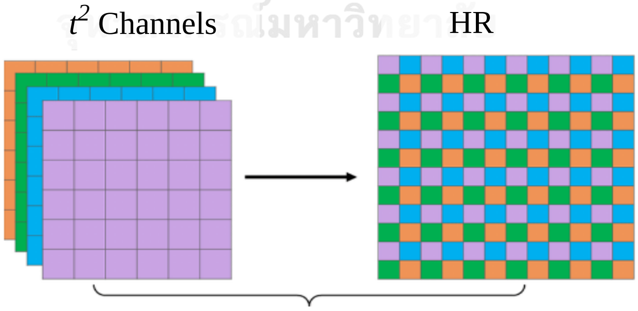


Figure 49: Pixel Shuffle

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSIONS

In this chapter, the model estimates the results of well-known parameters of PSNR and SSIM which are compared with the baseline method.

4.1 Experimental Setup

Initially, 15,000 iterations are performed to compare our method with the baseline method [19]. Bicubic downscale interpolation is applied to get LR frames. These are randomly cropped with a patch size of 32×32. We use k_1=5 and k_2 =40, residual blocks are utilized in the extracting features and for the construction module of HR frames, respectively. Data augmentation is performed by horizontal flipping and randomly rotating 90◦, 180◦, and 270◦. To employ the deformable alignment Pyramid, Cascading and Deformable (PCD) structure is adopted as in [64]. Adam optimizer [91] is applied with β_1 = 0.9 and $\beta_1 = 0.99$, where the cosine annealing gradually reduces the learning rate frame 4e-4 to 1e-7 for each batch. The batch size is set to be 8 and trained on the GPU of Nvidia Titan XP.

4.2 Evaluation metrics

4.2.1 Peak signal-to-noise ratio (PSNR)

The peak signal-to-noise ratio (PSNR) compares the peak signal to the corrupting noise. Equations 4.1 and 4.2 show the calculation in detail. Where n is the image's width, m is its height, $x(i,j)$ and $y(i,j)$ are its high-resolution and low-resolution pixels, respectively.

$$PSNR = 10log\frac{255}{MSE} \tag{4.1}$$

$$MSE = \frac{\Sigma_{i=0}^{n}\Sigma_{j=0}^{m}(x(i,j)-y(i,j))^2}{n\times m} \tag{4.2}$$

4.2.2 Structural similarity index measure (SSIM)

The structural similarity index measure (SSIM), is a technique that assesses the resemblance between the two pictures. The index can demonstrate how well the

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

prediction's output image performs compared to the reference image. The equation is as follows.

$$SSIM\,(x,y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{4.3}$$

Where:

- μ_x is the average of x, μ_y is the average of y,
- σ_x is the variance of x, σ_y is the variance of y,
- σ_{xy} is the covariance of x and y,
- c_1 and c_2 are constants needed to keep the formula valid and prevent the denominator from being zero.

4.3 Datasets

- For training, Vimeo-90K is used. These training sets include more than 60,000 7-frame video sequences training. The dataset has been used extensively in earlier VFI and VSR investigations.

Figure 50: Example images from Vimeo 90K dataset

- For testing, SPMC and Vid4 test sets are utilized.

→ SPMC consists of 30 different videos, each of them contains 31 frames.

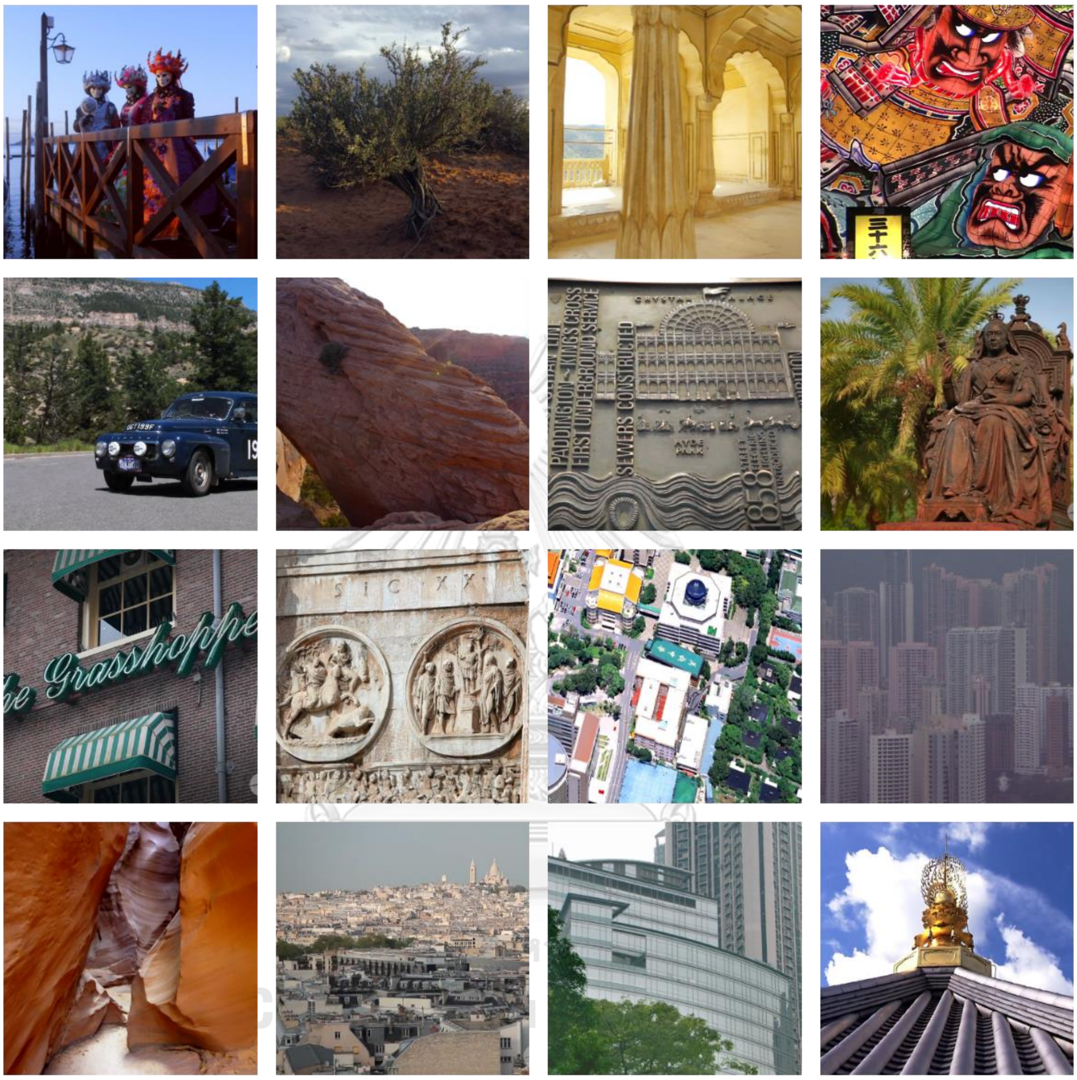


Figure 51: Example images from SPMC dataset

→ Vid4 consists of four sequences, walks (740×480, 47 frames), foliage (740×480, 49 frames), city (704×576, 34 frames), and calendar (720×576, 41 frames).

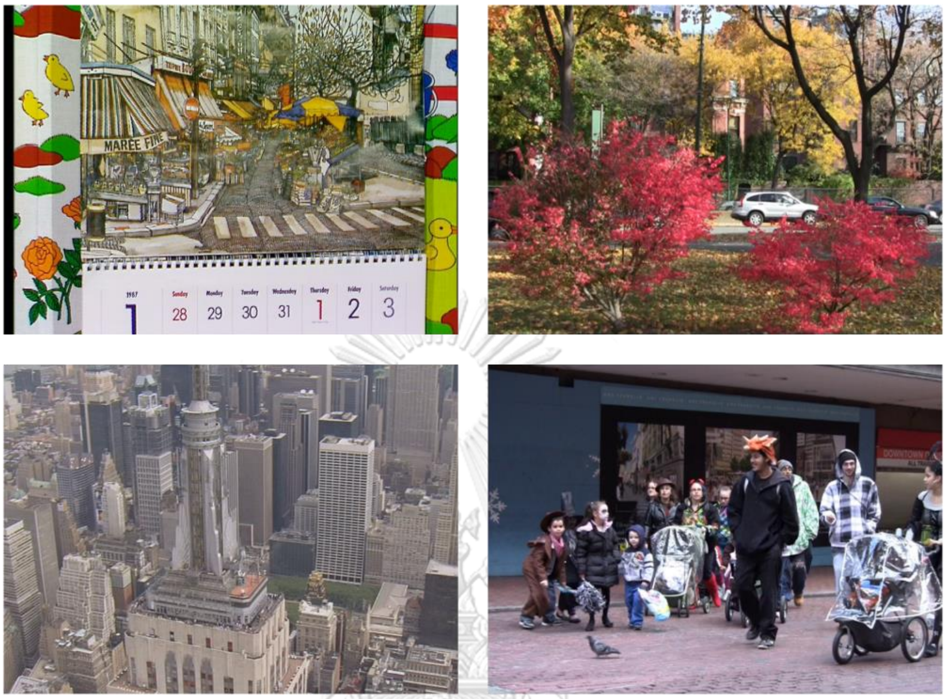


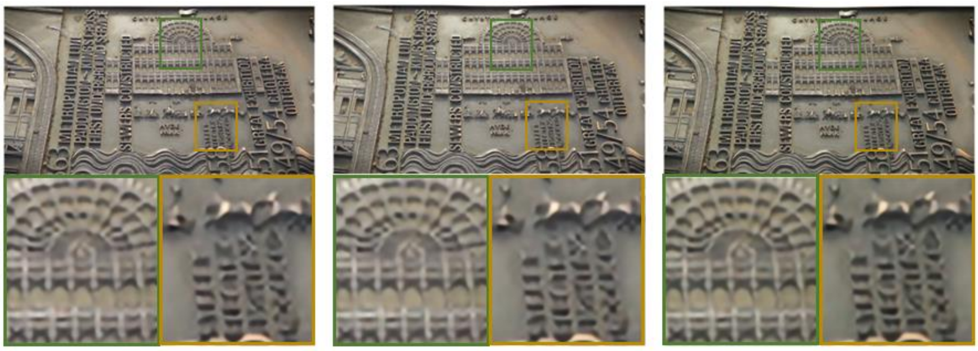
Figure 52: Example images from Vid4 dataset

4.4 Evaluation

The well-known parameter's structure similarity index (SSIM) and peak signal-to-noise ratio (PSNR) is assumed to estimate STVSR performance. To estimate the effectiveness of numerous networks, we evaluate the model sizes and interpretation time of the Vid4 dataset determined on the GPU of Nvidia Titan XP. Figure 54 to Figure 59 are the visual comparisons of frames of different images taken from the Vid4 dataset with the scale ×4. Figure 60, Figure 61, and Figure 62 are the visual comparisons of frames of different images taken from the Vid4 dataset with the scale ×4.

The following figures show the comparison of each iteration between the baseline and our method. According to this, the value of PSNR and SSIM are increasing with the increase in the number of iterations.

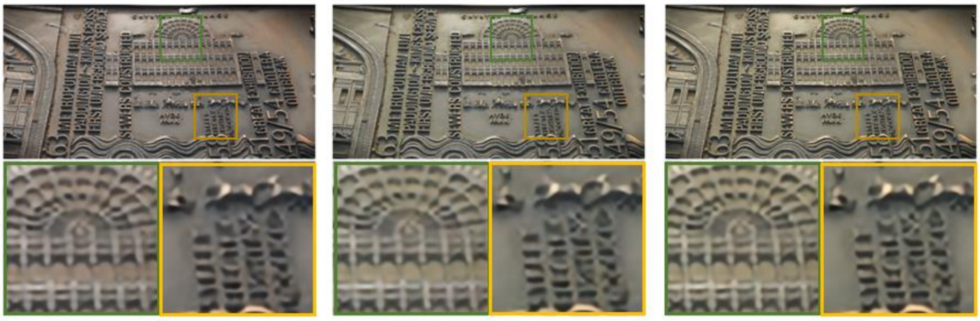
จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY



Zoom 5000_G
20.41/0.6230

Zoom 10000_G
20.78/0.6257

Zoom 15000_G
20.93/0.6250

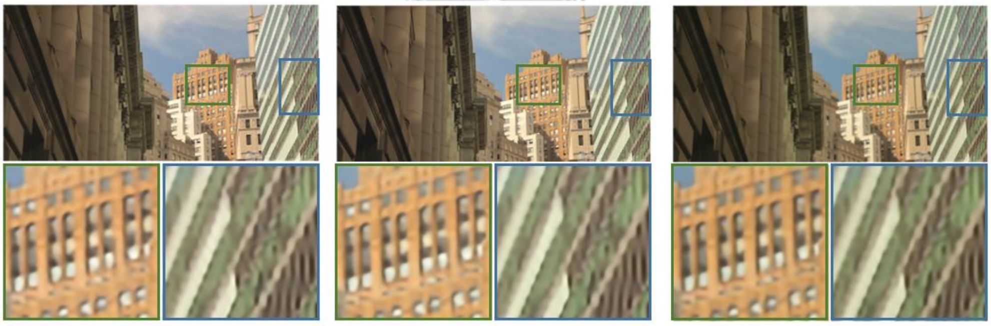


MHA 5000_G
20.77/0.6256

MHA 10000_G
20.95/0.6262

MHA 15000_G
21.07/0.6267

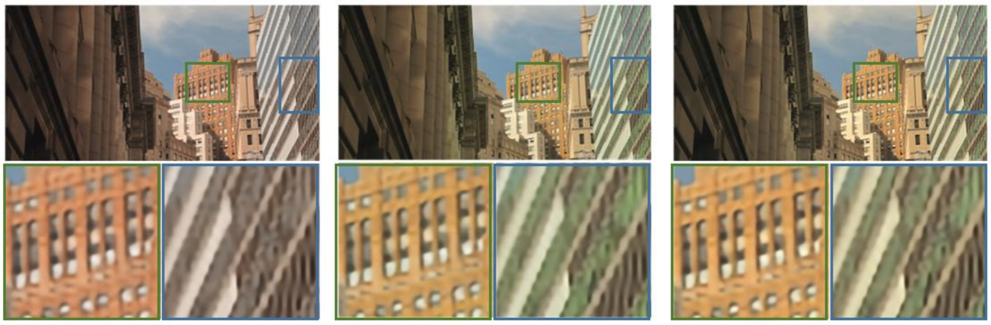
Figure 53: Visual comparisons of frames of "LDVTG_009" from SPMC dataset on scale ×4



Zoom 5000_G
22.76/0.7282

Zoom 10000_G
24.55/0.7373

Zoom 15000_G
24.73/0.7370

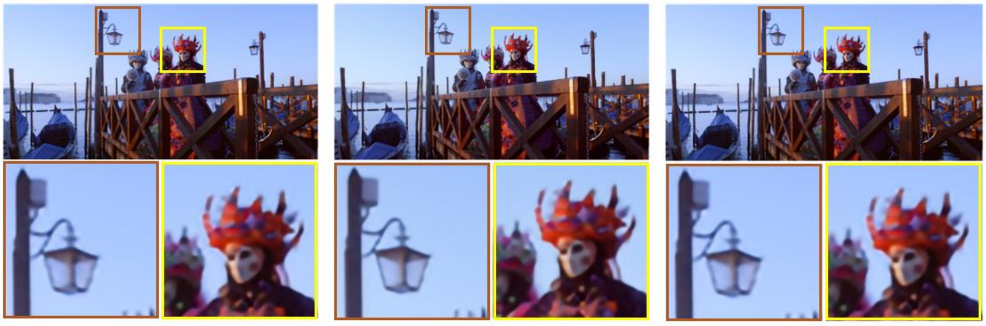


MHA 5000_G
23.64/0.7370

MHA 10000_G
24.67/0.7352

MHA 15000_G
24.74/0.7367

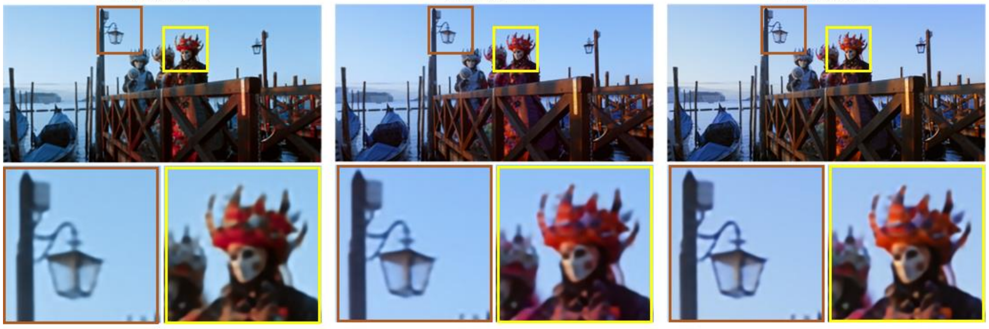
Figure 54: Visual comparisons of frames of "NYVTG_006" from SPMC dataset on scale ×4



Zoom 5000_G
24.52/0.8347

Zoom 10000_G
24.84/0.8329

Zoom 15000_G
24.90/0.8323

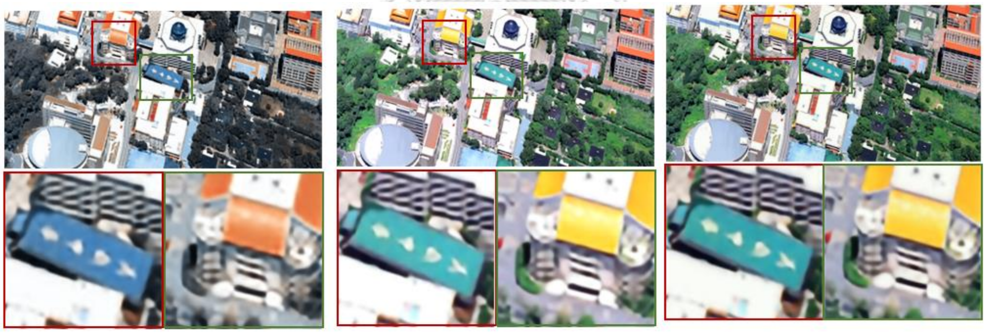


MHA 5000_G
24.59/0.8326

MHA 10000_G
24.87/0.8341

MHA 15000_G
25.00/0.8358

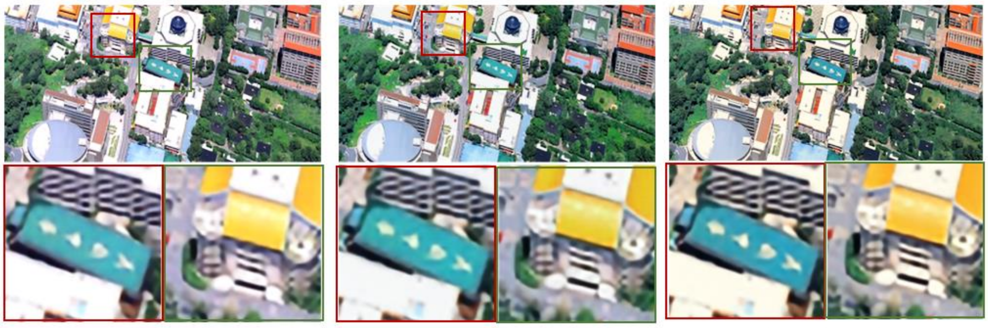
Figure 55: Visual comparisons of frames of “Veni3_011” from SPMC dataset on scale ×4



Zoom 5000_G
18.35/0.5172

Zoom 10000_G
19.09/0.5312

Zoom 15000_G
19.25/0.5292

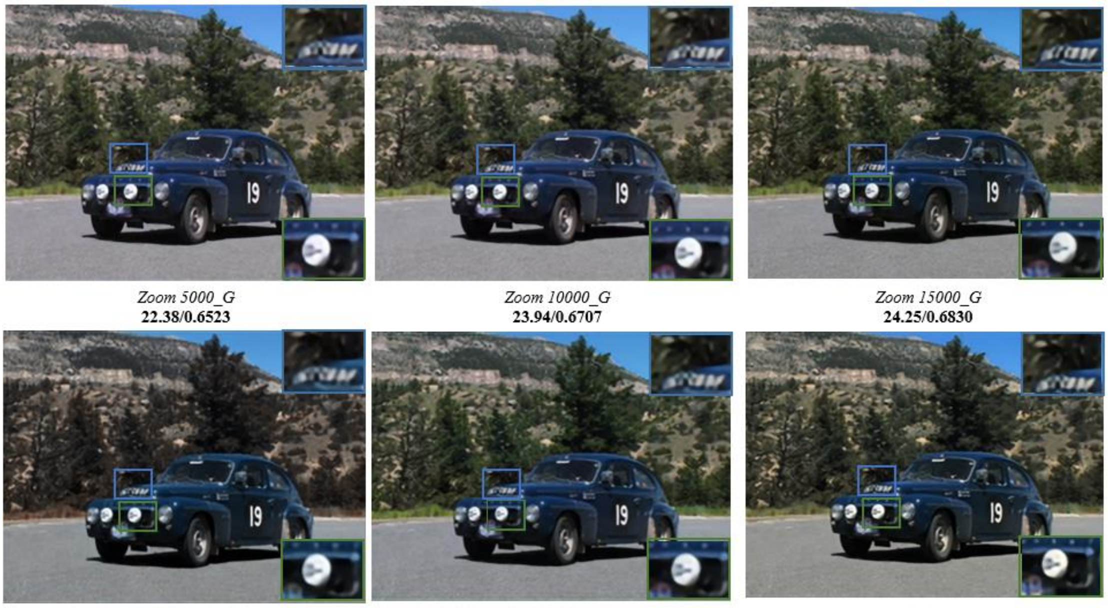


MHA 5000_G
19.10/0.5308

MHA 10000_G
19.29/0.5307

MHA 15000_G
19.36/0.5332

Figure 56: Visual comparisons of frames of “hdclub_001” from SPMC dataset on scale ×4



Zoom 5000_G
22.38/0.6523

Zoom 10000_G
23.94/0.6707

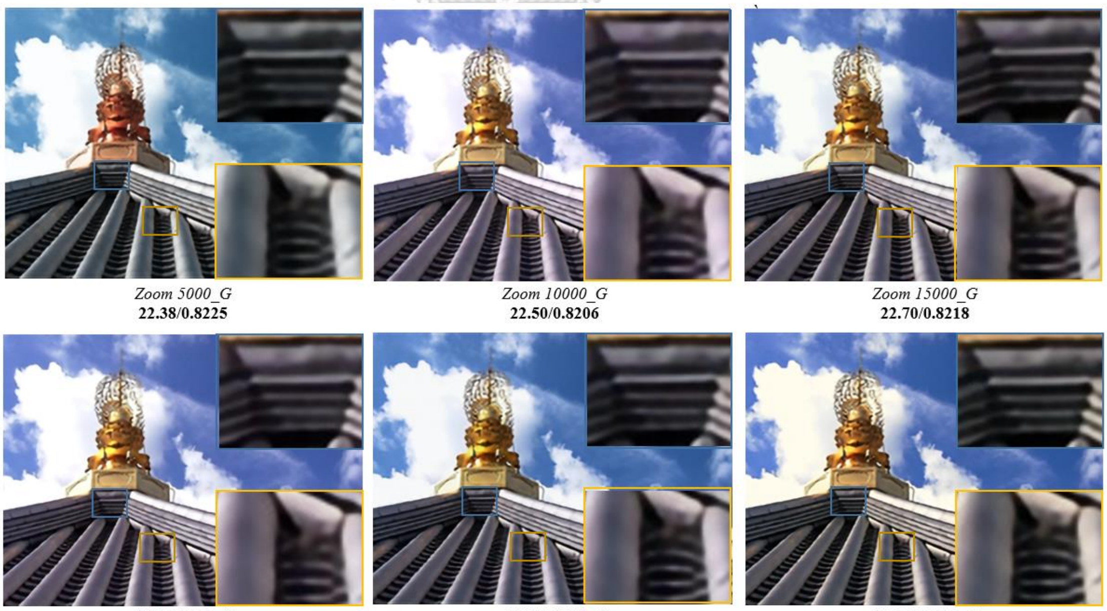
Zoom 15000_G
24.25/0.6830

MHA 5000_G
23.91/0.6696

MHA 10000_G
24.08/0.6927

MHA 15000_G
24.28/0.6916

Figure 57: Visual comparisons of frames of “car05” from SPMC dataset on scale ×4



Zoom 5000_G
22.38/0.8225

Zoom 10000_G
22.50/0.8206

Zoom 15000_G
22.70/0.8218

MHA 5000_G
22.46/0.8213

MHA 10000_G
22.68/0.8215

MHA 15000_G
22.74/0.8206

Figure 58: Visual comparisons of frames of “jvc_004” from SPMC dataset on scale ×4

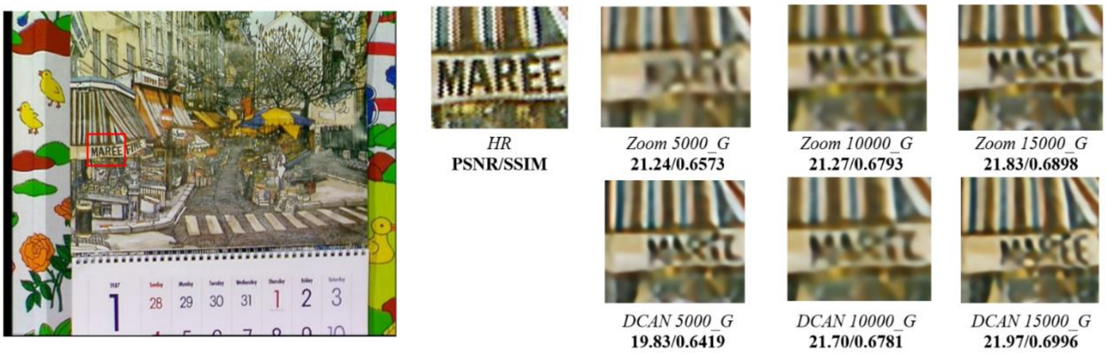


Figure 59: Visual comparisons of frames of “calendar” from Vid4 dataset on scale ×4

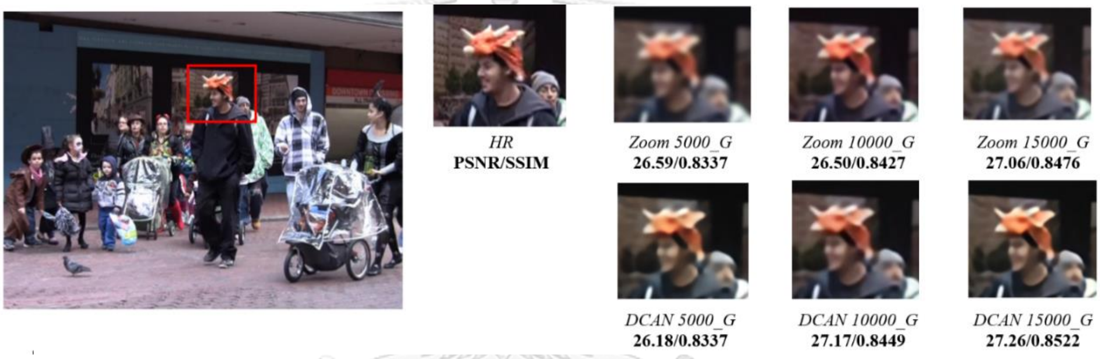


Figure 60: Visual comparisons of frames of “walk” from Vid4 dataset on scale ×4

จุฬาลงกรณ์มหาวิทยาลัย

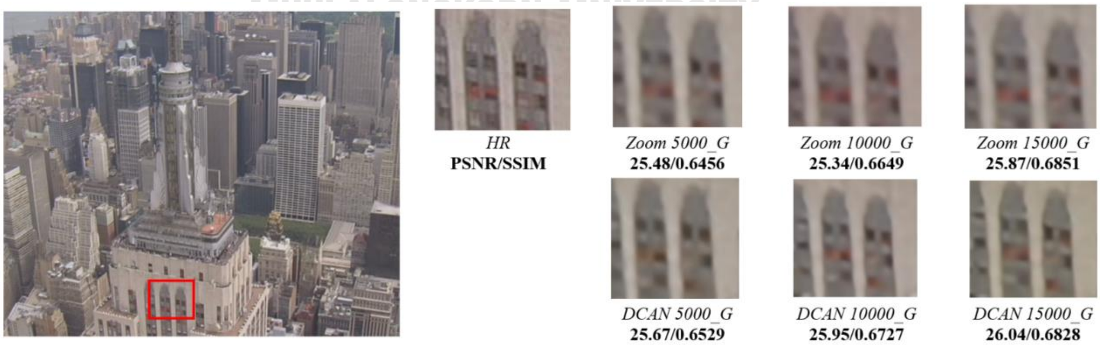
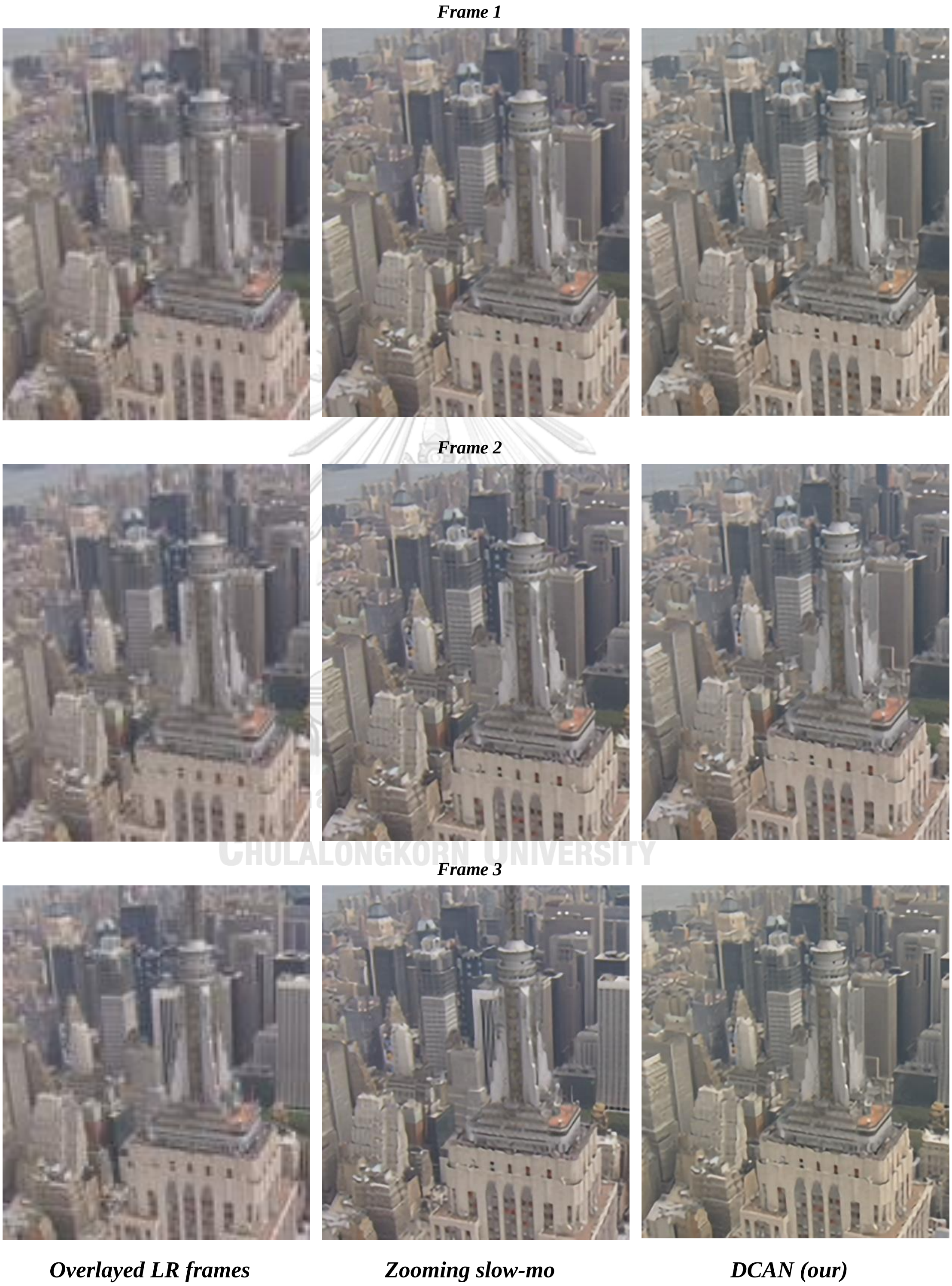


Figure 61: Visual comparisons of frames of “city” from Vid4 dataset on scale ×4

The visual comparison of overlayed LR video frames of two videos between the DCAN, our method, and the baseline is illustrated in Figure 63 and Figure 64, respectively. As seen in detail, after applying the different models, the LR frames give blurry artifacts in the baseline case



CHULALONGKORN UNIVERSITY

Figure 62 Visual comparison of video frames of “city”

However, our method outperforms in this regard and can reconstruct more visually appealing HR video frames with more accurate image structures and fewer blurring artifacts.

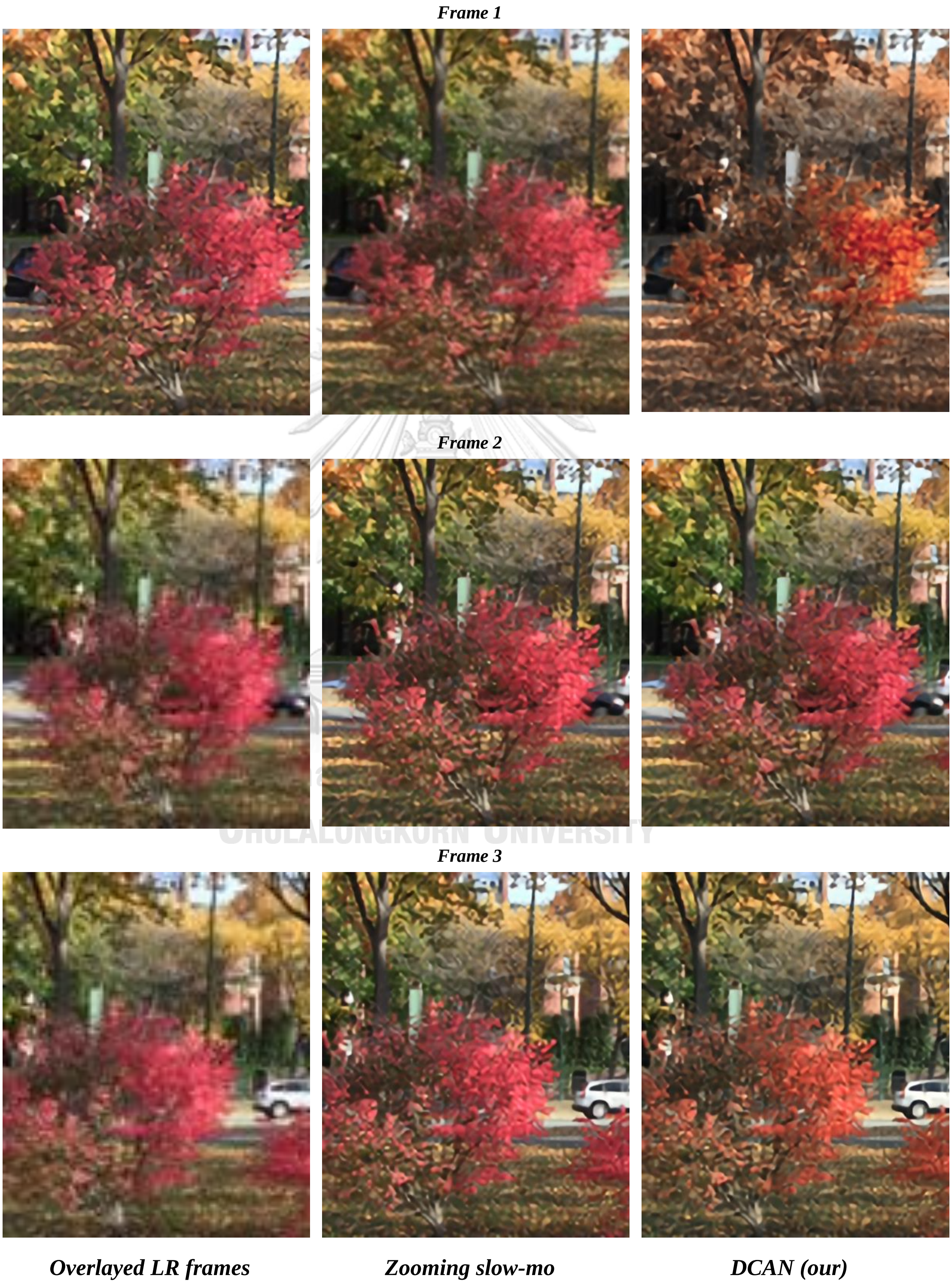


Figure 63 Visual comparison of video frames of “foliage”

CHULALONGKORN UNIVERSITY

The experiments show the results have good progress with our method of DCAN as shown in Table 1 and Table 2.

Table 1: Comparison of the proposed method on Vid4 dataset

Models	**Zooming Slow-Mo**		**DCAN (Ours)**	
Iterations	PSNR	SSIM	PSNR	SSIM
Vid4_5000_G	24.31	0.6914	24.46	0.6986
Vid4_10000_G	24.28	0.7034	24.79	0.7096
Vid4_15000_G	24.76	0.7167	24.93	0.7231

Table 2: Comparison of the proposed method on SPMC dataset

Models	**Zooming Slow-Mo**		**DCAN (Ours)**	
Iterations	PSNR	SSIM	PSNR	SSIM
SPMC_5000_G	23.35	0.6665	23.43	0.6670
SPMC_10000_G	23.45	0.6671	23.50	0.6706
SPMC_15000_G	23.79	0.6674	23.97	0.6717

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

The baseline method is compared based on PSNR and SSIM with the proposed method at different iterations, i.e., 5000_G to 15000_G with two different datasets. As shown in the tables, the PSNR increases with the increase in iterations and outperforms 0.27 dB and 0.31dB for Vid4 and SPMC respectively, in average PSNR compared to the state-of-the-art baseline method.

CHAPTER 5

CONCLUSION

5.1 Conclusion

In terms of the well-known parameters of PSNR and SSIM, the methodology of our work described in this study outperforms. The space-time VSR consecutive deep attention network reconstructs video sequences with HR frames obtained from the LR frames. The deep consecutive channel attention module is added to get the desired outcomes. Especially the multi-head attention controls the information mixing between features of the frame. This leads to the creation of rich representations. The network thoroughly studied the intra-relatedness between the SR tasks. After employing a deep consecutive attention network, this method adaptively learns about the advantageous local and global contexts to resolve the problems with video motion. Testing with several iterations produces beneficial results and achieves the necessary effectiveness above the prior network. This method is skilled and equipped to manage the captivating motion in video sequences.

5.2 Future Work

Implement the high-quality dataset to further improve the performance of the proposed algorithm.

Implement the transformer-based technique to analyze the behavior of the model.

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

REFERENCES

1. Liu, Z., et al. *Video frame synthesis using deep voxel flow*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
2. Long, G., et al. *Learning image matching by simply watching video*. in *European Conference on Computer Vision*. 2016. Springer.
3. Nimisha, T. and A. Rajagopalan, *Blind super-resolution of faces for surveillance*, in *Deep Learning-Based Face Analytics*. 2021, Springer. p. 119-136.
4. Wu, D., et al., *Streaming video over the Internet: approaches and directions.* IEEE Transactions on circuits and systems for video technology, 2001. **11**(3): p. 282-300.
5. Shishikui, Y. *Quality-of-experience evaluation of 8K ultra-high-definition television*. in *2021 IEEE International Conference on Image Processing (ICIP)*. 2021. IEEE.
6. Tun, E.E., S. Aramvith, and T. Onoye, *Low complexity mode selection for H. 266/VVC intra coding.* ICT Express, 2022. **8**(1): p. 83-90.
7. Bross, B., et al., *Developments in international video coding standardization after avc, with an overview of versatile video coding (vvc).* Proceedings of the IEEE, 2021. **109**(9): p. 1463-1493.
8. Xu, Y., et al., *TE-SAGAN: An Improved Generative Adversarial Network for Remote Sensing Super-Resolution Images.* Remote Sensing, 2022. **14**(10): p. 2425.
9. Turletti, T. and C. Huitema, *Videoconferencing on the Internet.* IEEE/ACM Transactions on networking, 1996. **4**(3): p. 340-351.
10. Parihar, A.S., et al., *A comprehensive survey on video frame interpolation techniques.* The Visual Computer, 2022. **38**(1): p. 295-319.
11. Liu, H., et al., *Video super-resolution based on deep learning: a comprehensive survey.* Artificial Intelligence Review, 2022: p. 1-55.
12. Patti, A.J., M.I. Sezan, and A.M. Tekalp, *Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time.* IEEE transactions on image processing, 1997. **6**(8): p. 1064-1076.
13. Bascle, B., A. Blake, and A. Zisserman. *Motion deblurring and super-resolution from an image sequence*. in *European conference on computer vision*. 1996. Springer.
14. Shechtman, E., Y. Caspi, and M. Irani, *Space-time super-resolution.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. **27**(4): p. 531-545.
15. Lu, Q., N. Xu, and X. Fang, *Motion-compensated frame interpolation with multiframe-based occlusion handling.* Journal of Display Technology, 2015. **12**(1): p. 45-54.
16. Tian, Y., et al. *Tdan: Temporally-deformable alignment network for video super-resolution*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
17. Muhammad, W., S. Aramvith, and T. Onoye, *Multi-scale Xception based depthwise separable convolution for single image super-resolution.* Plos one, 2021. **16**(8): p. e0249278.

CHULALONGKORN UNIVERSITY

18. Ganokratanaa, T., S. Aramvith, and N. Sebe, *Unsupervised anomaly detection and localization based on deep spatiotemporal translation network.* IEEE Access, 2020. **8**: p. 50312-50329.
19. Xiang, X., et al. *Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
20. Haris, M., G. Shakhnarovich, and N. Ukita. *Space-time-aware multi-resolution video enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
21. Xu, G., et al. *Temporal modulation network for controllable space-time video super-resolution*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
22. Meyer, S., et al. *Phase-based frame interpolation for video*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
23. Liu, Y.-L., et al. *Deep video frame interpolation using cyclic frame generation*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019.
24. Xue, T., et al., *Video enhancement with task-oriented flow.* International Journal of Computer Vision, 2019. **127**(8): p. 1106-1125.
25. Jiang, H., et al. *Super slomo: High quality estimation of multiple intermediate frames for video interpolation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
26. Bao, W., et al. *Depth-aware video frame interpolation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
27. Wang, Y., et al., *Depth map enhancement based on color and depth consistency.* The visual computer, 2014. **30**(10): p. 1157-1168.
28. Niklaus, S. and F. Liu. *Context-aware synthesis for video frame interpolation*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
29. Fourure, D., et al., *Residual conv-deconv grid network for semantic segmentation.* arXiv preprint arXiv:1707.07958, 2017.
30. Meyer, S., et al. *Phasenet for video frame interpolation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
31. Niklaus, S., L. Mai, and F. Liu. *Video frame interpolation via adaptive convolution*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
32. Niklaus, S., L. Mai, and F. Liu. *Video frame interpolation via adaptive separable convolution*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
33. van Amersfoort, J., et al., *Frame interpolation with multi-scale deep loss functions and generative adversarial networks.* arXiv preprint arXiv:1711.06045, 2017.
34. Peleg, T., et al. *Im-net for high resolution video frame interpolation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
35. Vidanpathirana, M., et al., *Tracking and frame-rate enhancement for real-time 2D human pose estimation.* The Visual Computer, 2020. **36**(7): p. 1501-1519.
36. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in

Proceedings of the 22nd ACM international conference on Multimedia. 2014.

37. Ahn, H.-E., J. Jeong, and J.W. Kim, *A fast 4k video frame interpolation using a hybrid task-based convolutional neural network.* Symmetry, 2019. **11**(5): p. 619.
38. Baker, S., et al., *A database and evaluation methodology for optical flow.* International journal of computer vision, 2011. **92**(1): p. 1-31.
39. Herbst, E., S. Seitz, and S. Baker, *Occlusion reasoning for temporal interpolation using optical flow.* Department of Computer Science and Engineering, University of Washington, Tech. Rep. UW-CSE-09-08-01, 2009.
40. Soomro, K., A.R. Zamir, and M. Shah, *A dataset of 101 human action classes from videos in the wild.* Center for Research in Computer Vision, 2012. **2**(11).
41. Ancuti, C., et al., *Video enhancement using reference photographs.* The Visual Computer, 2008. **24**(7): p. 709-717.
42. Liu, H., et al., *Multiple hypotheses Bayesian frame rate up-conversion by adaptive fusion of motion-compensated interpolations.* IEEE transactions on circuits and systems for video technology, 2012. **22**(8): p. 1188-1198.
43. Yu, Z., et al., *Multi-level video frame interpolation: Exploiting the interaction among different levels.* IEEE Transactions on Circuits and Systems for Video Technology, 2013. **23**(7): p. 1235-1248.
44. Lee, W.H., K. Choi, and J.B. Ra, *Frame rate up conversion based on variational image fusion.* IEEE Transactions on image processing, 2013. **23**(1): p. 399-412.
45. Zhao, Y., G. Ge, and Q. Sun. *Frame rate up-conversion based on edge information*. in *2019 7th International Conference on Information, Communication and Networks (ICICN)*. 2019. IEEE.
46. Li, R., et al., *A Low-Complex Frame Rate Up-Conversion with Edge-Preserved Filtering.* Electronics, 2020. **9**(1): p. 156.
47. Lee, H., et al. *Adacof: Adaptive collaboration of flows for video frame interpolation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
48. Werlberger, M., et al. *Optical flow guided TV-L 1 video interpolation and restoration*. in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. 2011. Springer.
49. Cheng, X. and Z. Chen. *Video frame interpolation via deformable separable convolution*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.
50. Cheng, X. and Z. Chen, *Multiple video frame interpolation via enhanced deformable separable convolution.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
51. Shi, Z., et al., *Video frame interpolation via generalized deformable convolution.* IEEE Transactions on Multimedia, 2021. **24**: p. 426-439.
52. Ding, T., et al. *Cdfi: Compression-driven network design for frame interpolation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
53. Protter, M., et al., *Generalizing the nonlocal-means to super-resolution reconstruction.* IEEE Transactions on image processing, 2008. **18**(1): p. 36-51.
54. Takeda, H., et al., *Super-resolution without explicit subpixel motion estimation.* IEEE Transactions on Image Processing, 2009. **18**(9): p. 1958-1975.
55. Sun, D., et al. *Pwc-net: Cnns for optical flow using pyramid, warping, and cost*

volume. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

56. Yi, P., et al. *Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
57. Bao, W., et al., *Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement.* IEEE transactions on pattern analysis and machine intelligence, 2019. **43**(3): p. 933-948.
58. Liao, R., et al. *Video super-resolution via deep draft-ensemble learning*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
59. Sun, X., et al., *VSRNet: End-to-end video segment retrieval with text query.* Pattern Recognition, 2021. **119**: p. 108027.
60. Shi, W., et al. *Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
61. Tao, X., et al. *Detail-revealing deep video super-resolution*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
62. Sajjadi, M.S., R. Vemulapalli, and M. Brown. *Frame-recurrent video super-resolution*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
63. Dai, J., et al. *Deformable convolutional networks*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
64. Wang, X., et al. *Edvr: Video restoration with enhanced deformable convolutional networks*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
65. Zhu, X., et al. *Deformable convnets v2: More deformable, better results*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
66. Wang, H., et al., *Deformable non-local network for video super-resolution.* IEEE Access, 2019. **7**: p. 177734-177744.
67. Wang, X., et al. *Non-local neural networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
68. Hui, Z., et al., *Progressive perception-oriented network for single image super-resolution.* Information Sciences, 2021. **546**: p. 769-786.
69. Ying, X., et al., *Deformable 3d convolution for video super-resolution.* IEEE Signal Processing Letters, 2020. **27**: p. 1500-1504.
70. Chen, J., et al., *Vesr-net: The winning solution to youku video enhancement and super-resolution challenge.* arXiv preprint arXiv:2003.02115, 2020.
71. Yan, B., C. Lin, and W. Tan. *Frame and feature-context video super-resolution.* in *Proceedings of the AAAI conference on artificial intelligence*. 2019.
72. Lucas, A., et al., *Generative adversarial networks and perceptual losses for video super-resolution.* IEEE Transactions on Image Processing, 2019. **28**(7): p. 3312-3327.
73. Tran, D., et al. *Learning spatiotemporal features with 3d convolutional networks*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
74. Jo, Y., et al. *Deep video super-resolution network using dynamic upsampling*

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

filters without explicit motion compensation. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

75. Haris, M., G. Shakhnarovich, and N. Ukita. *Recurrent back-projection network for video super-resolution*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
76. Shi, X., et al., *Convolutional LSTM network: A machine learning approach for precipitation nowcasting.* Advances in neural information processing systems, 2015. **28**.
77. Itti, L., C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis.* IEEE Transactions on pattern analysis and machine intelligence, 1998. **20**(11): p. 1254-1259.
78. Rensink, R.A., *The dynamic representation of scenes.* Visual cognition, 2000. **7**(1-3): p. 17-42.
79. Larochelle, H. and G.E. Hinton, *Learning to combine foveal glimpses with a third-order Boltzmann machine.* Advances in neural information processing systems, 2010. **23**.
80. Wang, F., et al. *Residual attention network for image classification*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
81. Hu, J., L. Shen, and G. Sun. *Squeeze-and-excitation networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
82. Zeiler, M.D. and R. Fergus. *Visualizing and understanding convolutional networks*. in *European conference on computer vision*. 2014. Springer.
83. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
84. Shechtman, E., Y. Caspi, and M. Irani. *Increasing space-time resolution in video*. in *European Conference on Computer Vision*. 2002. Springer.
85. Mudenagudi, U., S. Banerjee, and P.K. Kalra, *Space-time super-resolution using graph-cut optimization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. **33**(5): p. 995-1008.
86. Boykov, Y., O. Veksler, and R. Zabih, *Fast approximate energy minimization via graph cuts.* IEEE Transactions on pattern analysis and machine intelligence, 2001. **23**(11): p. 1222-1239.
87. Geman, S. and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.* IEEE Transactions on pattern analysis and machine intelligence, 1984(6): p. 721-741.
88. Kim, S.Y., J. Oh, and M. Kim. *Fisr: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.
89. Zhang, Y., et al. *Image super-resolution using very deep residual channel attention networks*. in *Proceedings of the European conference on computer vision (ECCV)*. 2018.
90. Schuster, M. and K.K. Paliwal, *Bidirectional recurrent neural networks.* IEEE transactions on Signal Processing, 1997. **45**(11): p. 2673-2681.
91. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

VITA

NAME	Talha Saleem
DATE OF BIRTH	19 Nov 1994
PLACE OF BIRTH	Faisalabad, Pakistan.
INSTITUTIONS ATTENDED	Chulalongkorn University (2020-Continue) The University of Faisalabad (2014-2018)
HOME ADDRESS	29/55, Siam Condominium, 2 Rama IX 3 Alley, Huai Khwang, Bangok,10310.
PUBLICATION	DCAN: Deep Consecutive Attention Network for Video Super-Resolution By Talha Saleem, Sovann Chen, Supavadee Aramvith Conference : Asia Pacific Signal and Information Processing Association Annual Summit and Conference 2022

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY