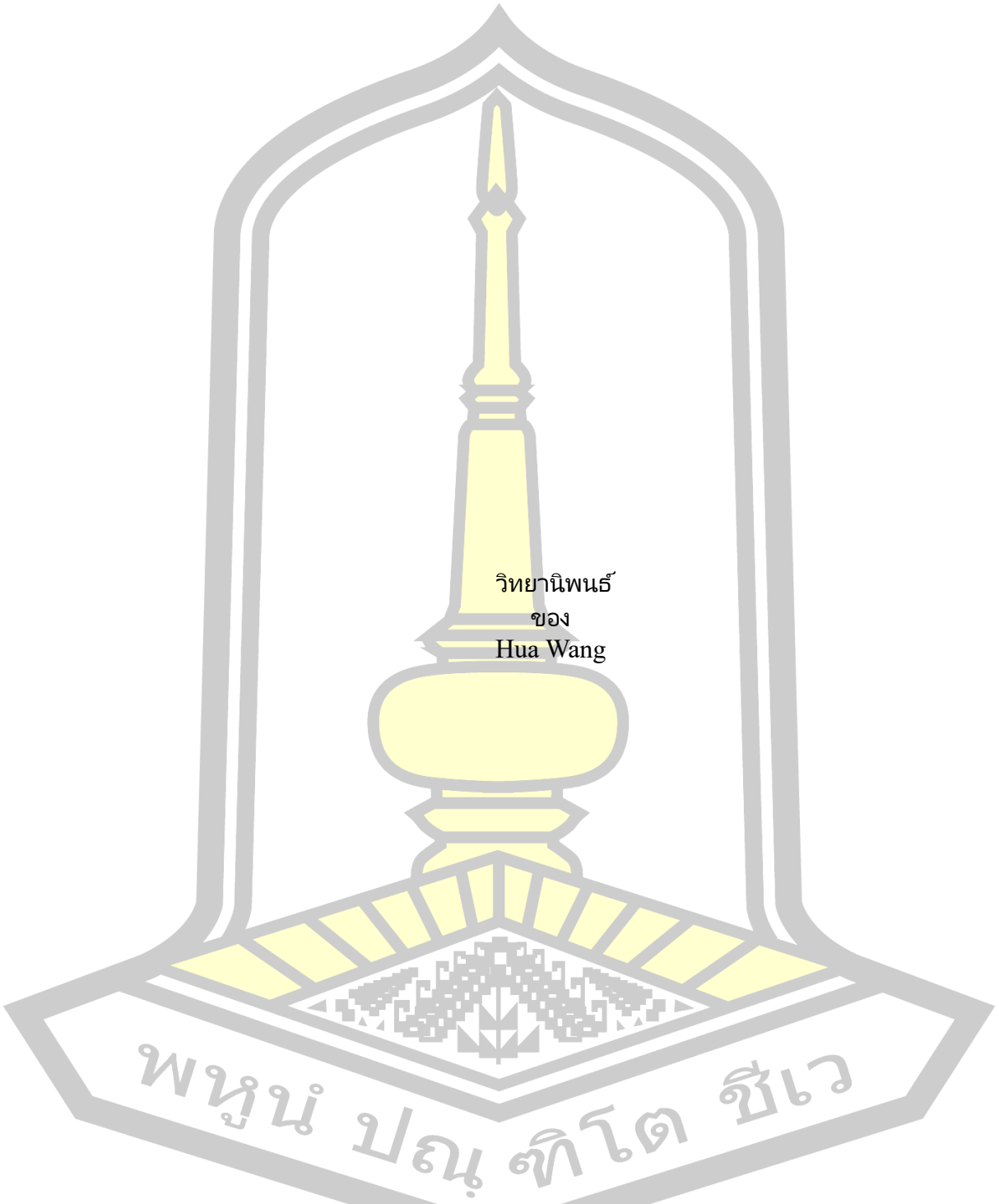
Efficient Networks for Video Quality Enhancement

Hua Wang

A Thesis Submitted in Partial Fulfillment of Requirements for
degree of Doctor of Philosophy in Computer Science
March 2025

Efficient Networks for Video Quality Enhancement



วิทยานิพนธ์
ของ
Hua Wang

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
มีนาคม 2568

Efficient Networks for Video Quality Enhancement

Hua Wang

A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Computer Science)

March 2025

The examining committee has unanimously approved this Thesis, submitted by Mr. Hua Wang , as a partial fulfillment of the requirements for the Doctor of Philosophy Computer Science at Mahasarakham University

Examining Committee

.....Chairman
(Assoc. Prof. Suphakant Phimoltares , Ph.D.)

.....Advisor
(Asst. Prof. Rapeeporn Chamchong , Ph.D.)

.....Co-advisor
(Pornntiwa Pawara , Ph.D.)

.....Committee
(Asst. Prof. Chatklaw Jareanpon , Ph.D.)

.....Committee
(Assoc. Prof. Olarik Surinta , Ph.D.)

Mahasarakham University has granted approval to accept this Thesis as a partial fulfillment of the requirements for the Doctor of Philosophy Computer Science

.....
(Assoc. Prof. Jantima Polpinij , Ph.D.)
Dean of The Faculty of Informatics

.....
(Prof. Anongrit Kangrang , Ph.D.)
Acting Dean of Graduate School

TITLE Efficient Networks for Video Quality Enhancement

AUTHOR Hua Wang

ADVISORS Assistant Professor Rapeeporn Chamchong , Ph.D.
Pornntiwa Pawara , Ph.D.

DEGREE Doctor of Philosophy **MAJOR** Computer Science

UNIVERSITY Mahasarakham University **YEAR** 2025

ABSTRACT

Video restoration has become increasingly important with the growing demand for high-quality video content across various applications. This thesis addresses two fundamental challenges in video restoration: space-time video super-resolution and video deblurring. For space-time video super-resolution, we propose a novel deformable attention network (DANet) that effectively handles both spatial and temporal super-resolution in a unified framework. The network features a deformable interpolation block for accurate frame synthesis and a temporal fusion module for efficient multi-frame information utilization. For video deblurring, we develop a wavelet-based blur-aware decoupled network (WBDNet) that innovatively decomposes the deblurring task into structure recovery and detail enhancement through wavelet transform. The network employs a multi-scale progressive fusion module for structural reconstruction and a blur-aware detail enhancement module that leverages sharpness priors for refined detail restoration. Extensive experiments on multiple benchmark datasets demonstrate that our proposed methods achieve superior performance compared to state-of-the-art approaches in terms of both objective metrics and visual quality, while maintaining reasonable computational efficiency. The methods developed in this thesis advance the field of video restoration and show strong practical value for applications ranging from multimedia entertainment to surveillance systems.

Keyword : Video Restoration; Space-Time Video Super-Resolution; Video Deblurring; Deep Learning; Deformable Convolution; Attention Mechanism; Wavelet Transform

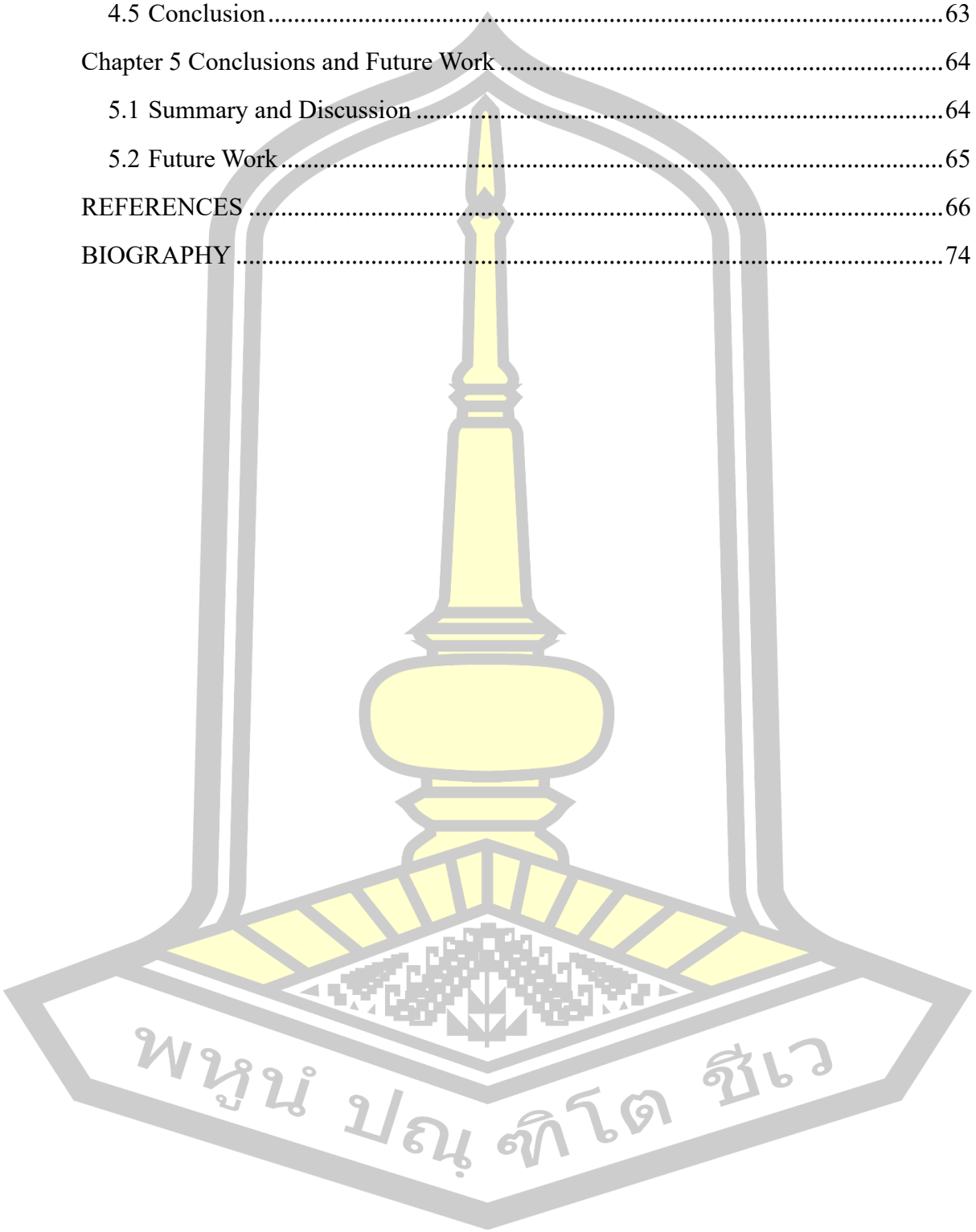
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have supported me throughout my doctoral studies. First and foremost, I am deeply grateful to my former supervisor, Prof. Phatthanaphong Chompoowises, and my current supervisor, Prof. Rapeeporn Chamchong, for their meticulous guidance and invaluable support. They have not only provided sufficient experimental resources but also offered valuable advice throughout my research and thesis writing. During my time in Thailand, I have been deeply fascinated by its unique cultural traditions and impressed by the warmth of the local people. I feel truly fortunate to have completed my doctoral studies in this beautiful country - an experience that has enriched both my academic horizons and personal life. Finally, I would like to acknowledge the financial support from the Faculty of Informatics, Mahasarakham University, which has been crucial in enabling me to successfully complete this thesis.

Hua Wang

TABLE OF CONTENTS

Page



LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

Introduction

1.1 Background

In the era of ubiquitous video applications, the demand for high-quality video content has grown dramatically. However, real-world video data often suffers from multiple degradation factors that compromise visual quality. Two critical challenges stand out: low spatial-temporal resolution caused by historical device limitations or compression artifacts, and motion blur resulting from camera shake, fast-moving objects, or environmental disturbances. Addressing these issues is essential for applications ranging from surveillance to consumer video enhancement and augmented reality.

Video super-resolution (VSR) aims to reconstruct high-resolution (HR) videos from low-resolution (LR) inputs by recovering spatial details and temporal continuity. Early VSR methods primarily focused on spatial reconstruction [1-3], leveraging multi-frame alignment or deep learning architectures. However, modern applications require simultaneous enhancement of both spatial resolution and temporal frame rate—a task termed Space-Time Video Super-Resolution (STVSR). While existing methods have achieved remarkable progress, they often face computational bottlenecks due to heavy architectures [4, 5], limiting their practicality in real-time scenarios such as live streaming or video conferencing.

Video deblurring addresses another pervasive challenge: restoring sharp details from motion-blurred frames. Blur arises from dynamic scenes, handheld camera usage, or low-light conditions, distorting structural integrity and erasing high-frequency textures. Traditional deblurring approaches relied on explicit motion modeling (e.g., optical flow) [6-8], but they struggled with error propagation in complex scenarios. Recent advances in deformable convolution [9-11] and transformer architectures [12-14] have improved alignment accuracy, yet balancing structural recovery and detail enhancement remains an open problem.

These two tasks, VSR and deblurring, are deeply interconnected. Low-quality videos often exhibit both resolution limitations and motion blur, necessitating joint or sequential restoration. This thesis proposes novel frameworks for efficient video super-resolution and blur-aware video deblurring, advancing the field toward holistic video restoration.

1.2 Objective of Research

The overarching goal of this study is to advance video restoration through two directions: (1) developing a space-time video super-resolution (STVSR) framework that optimizes the trade-off between computational efficiency and reconstruction quality, and (2) designing a blur-aware video deblurring network that decouples structural recovery from detail enhancement using frequency-domain priors.

For STVSR, the primary objective is to overcome the limitations of existing alignment-based methods by introducing deformable convolution with hierarchical feature fusion. This approach aims to robustly handle multi-scale motions while minimizing redundant computations. A secondary objective involves designing lightweight temporal fusion modules to improve computational efficiency without sacrificing perceptual quality—critical for applications like live broadcasting and video conferencing.

In video deblurring, the research focuses on disentangling structural and textural restoration through wavelet-based frequency decomposition. A key objective is to leverage sharpness priors from motion estimation to guide attention mechanisms, allowing the network to selectively enhance features in less blurred regions. Additionally, the work explores hybrid alignment strategies combining optical flow for coarse structure correction and deformable convolution for refined detail alignment.

1.3 Contribution of Research

This research advances the field of video restoration through two frameworks, each addressing critical challenges in video super-resolution and deblurring. The primary contributions are as follows:

1) Deformable Attention Network for Space-Time Video Super-Resolution

- We design a deformable interpolation block that enhances the capability of deformable convolution in handling large-scale and complex motions, enabling more accurate intermediate frame generation.
- We propose a temporal feature shuffle block that facilitates complementary temporal information learning across multiple frame features, and develop a motion feature enhancement block that selectively emphasizes motion-related features to optimize temporal information exchange.
- Experimental results on multiple datasets demonstrate the effectiveness and superiority of the proposed method, showing significant advantages over state-

of-the-art methods.

2) Wavelet-Based Blur-Aware Decoupled Network for Video Deblurring

- We propose a novel multi-scale progressive fusion (MSPF) module that effectively reconstructs structural information through multi-scale feature fusion, significantly improving the restoration of low-frequency components.
- We design an innovative blur-aware detail enhancement module (BADE) that adaptively perceives and extracts sharp features across multiple frames for precise detail recovery, enabling more effective high-frequency information restoration.
- Experimental results on benchmark datasets demonstrate that our proposed method achieves significant performance improvements over state-of-the-art approaches, particularly in preserving both structural integrity and detail fidelity.

1.4 Scope of Research

The research scope spans two core tasks:

1) Video Super-Resolution: Focused on STVSR, we address LR videos with low frame rates. The goal is joint spatial upscaling ($4\times$) and temporal interpolation ($2\times$) under a unified framework. The primary datasets include Vimeo-90K for training and Vid4/SPMCS/REDS for evaluation, with metrics emphasizing peak signal to noise ratio (PSNR), structural similarity (SSIM) and model efficiency.

2) Video Deblurring: Targeting dynamic blur from camera shake and object motion, experiments cover synthetic (DVD, GoPro) and real-world (BSD) datasets. Evaluations measure PSNR, SSIM, structural integrity and detail fidelity.

1.5 Thesis Organization

The remainder of this thesis is structured as follows:

Chapter 2 reviews related work in video restoration, covering spatial super-resolution, temporal interpolation, and deblurring methods. It establishes the theoretical foundations including residual learning, attention mechanisms and deformable convolution. The chapter also introduces common evaluation metrics used in video restoration tasks.

Chapter 3 presents a novel deformable attention network (DANet) for space-time video super-resolution. The network architecture consists of three main modules: an interpolation module featuring deformable interpolation blocks for accurate frame synthesis, a temporal fusion module with feature shuffle and motion enhancement

mechanisms for effective temporal information utilization, and a reconstruction module for final high-resolution frame generation. Comprehensive experiments demonstrate the superior performance of DANet in terms of both restoration quality and computational efficiency.

Chapter 4 proposes a wavelet-based blur-aware decoupled network (WBDNet) for video deblurring. The network decouples the deblurring task into structure recovery and detail enhancement through wavelet transform. A multi-scale progressive fusion module handles structural reconstruction while a blur-aware detail enhancement module leverages sharpness priors for refined detail restoration. Extensive evaluations validate the effectiveness of this decoupled approach across various blur scenarios.

Chapter 5 summarizes the key contributions and findings of this research, and suggests promising directions for future work in video restoration.

1.6 Definition of Terms

Table 1.1 Terminology

Term	Definition
Video Super-Resolution (VSR)	The process of reconstructing high-resolution video frames from their low-resolution counterparts.
Video Frame Interpolation (VFI)	The technique of generating intermediate frames between existing video frames to increase temporal resolution.
Space-Time Video Super-Resolution (STVSR)	A joint task that simultaneously enhances both spatial resolution and temporal frame rate of videos.
Video Deblurring	The process of removing motion blur and recovering sharp details from blurred video frames.
Deformable Convolution	An extension of regular convolution that allows for adaptive sampling locations, making it particularly effective for handling geometric transformations.
Attention Mechanism	A technique that enables networks to focus on relevant features while suppressing less important ones, improving information processing efficiency.
Wavelet Transform	A mathematical tool that decomposes signals into different frequency components, allowing separate processing of structural and detail information.
Optical Flow	A technique for estimating pixel-wise motion between video frames, commonly used for frame warping.

Chapter 2

Literature Review

This chapter systematically reviews the foundational theories and methodological advancements in video restoration, focusing on three main aspects: video resolution enhancement, video deblurring, and their evaluation metrics. Video restoration has become increasingly important in various applications, from entertainment to surveillance and video streaming. The continuous advancement of deep learning has dramatically transformed this field, introducing novel architectures and methodologies that significantly improve restoration quality.

2.1 Theoretical Basis

2.1.1 Residual Learning

He et al. proposed ResNet [15], which alleviates the gradient disappearance problem when the network depth is too deep by stacking residual blocks and has shown excellent performance. Ledig et al. first introduced this structure to the super-resolution task and proposed SRResNet [16].

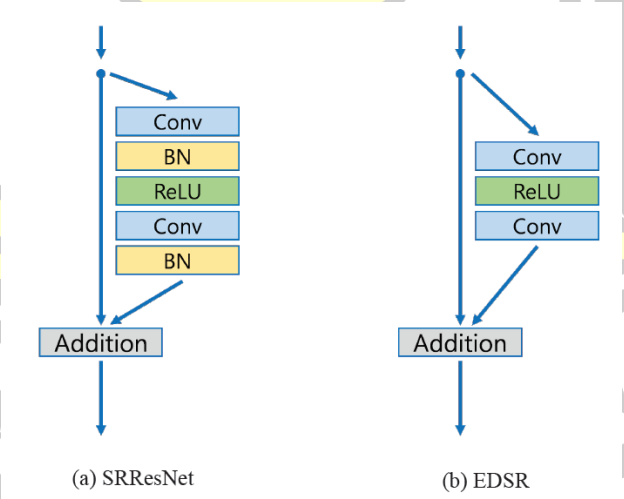


Figure 2.1 Schematic diagram of residual block structure [17].

However, the original ResNet is used to solve high-level visual problems, such as image classification, object detection, and other tasks. It is not optimal to apply it directly to super-resolution tasks without any changes. For this reason, Lim et al. proposed EDSR [17]. They found through analysis that the normalization layer in the

residual block would make the network lose range flexibility and discard the original contrast information of the image, so the normalization layer was removed and the structure of the residual block was simplified, as shown in Figure 2.1. The architecture of EDSR can be seen in Figure 2.2. EDSR starts with a 3×3 convolutional layer to extract features from low-resolution images. Then it applies multiple residual blocks to learn high-resolution residual features from the low-resolution ones. Finally, it uses a sub-pixel convolutional layer to scale up the feature map, and outputs a high-resolution image through a convolutional layer.

The experimental results of EDSR show that the proposed improvement of residual block structure is effective for super-resolution task, and it wins the first prize in the NTIRE2017 competition. With the development of image restoration research, this kind of residual block is widely used in the process of image feature extraction and shows good performance and efficiency.

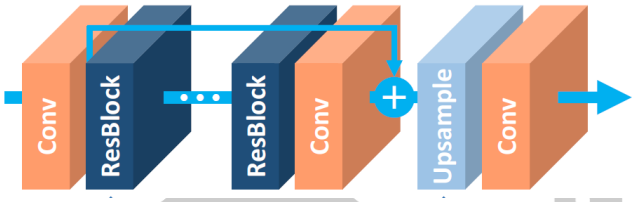


Figure 2.2 The network architecture of EDSR [17].

2.1.2 Attention Mechanism

The Attention mechanisms, inspired by human cognitive processes, enable neural networks to selectively focus on the most informative components of input data. In video restoration, these mechanisms have become fundamental building blocks that dynamically modulate feature representations across multiple dimensions, allowing networks to emphasize crucial information while suppressing less relevant features.

The core principle of attention mechanisms involves computing attention weights through trainable parameters, typically normalized by sigmoid or softmax functions to generate scaling factors between 0 and 1. These weights act as adaptive gates that modulate input features through element-wise multiplication, effectively creating a learned importance map that guides the restoration process. Attention mechanisms in video restoration can be categorized into several key types based on their operational dimensions:

1) Channel Attention: Operating along the feature channel dimension, this mechanism learns to emphasize informative channels while suppressing less relevant

ones [18]. Recent works [19-21] have shown that channel attention can effectively capture inter-channel dependencies and feature relationships. For instance, the SE-like (Squeeze-and-Excitation) modules have been widely adopted to recalibrate channel-wise feature responses, leading to improved restoration of texture details and structural information.

2) Spatial Attention: These modules [22] learn to identify spatially important regions within frames, such as edges, textures, and motion boundaries. Advanced spatial attention designs incorporate deformable convolutions [23] to adaptively sample features from relevant spatial locations. This is particularly effective for handling local artifacts and preserving spatial details during restoration.

3) Temporal Attention: Unique to video processing, temporal attention mechanisms evaluate and weight the relevance of different frames or temporal features. These modules are crucial for handling temporal misalignment, occlusions, and varying motion patterns. Recent works [1, 24] have demonstrated that temporal attention can effectively aggregate information from multiple frames while maintaining temporal consistency.

4) Self-Attention and Multi-head Attention: Transformer-based architectures [9, 12, 14, 25] have revolutionized video restoration by introducing self-attention mechanisms that model global dependencies across all positions in space and time. Multi-head attention extends this by learning diverse relationship patterns simultaneously through parallel attention heads. This approach has proven particularly effective for capturing long-range dependencies and complex spatial-temporal correlations.

The evolution of attention mechanisms continues to drive progress in video restoration, offering increasingly sophisticated ways to model and utilize spatial-temporal dependencies. As efficiency optimization techniques mature, these mechanisms are becoming more practical for real-world applications while maintaining their powerful feature modeling capabilities.

2.1.3 Deformable Convolution

The deformable convolutional network was initially proposed by Dai et al. [26] and later enhanced by Zhu et al. [27] Unlike ordinary convolutional neural network (CNN) that utilize a fixed geometric structure in a layer, deformable convolution (DConv) can model geometric transformations. As illustrated in Figure 2.3, DConv adds extra offsets to the regular grid sampling positions of standard convolution, making the sampling grid arbitrarily deformable. The offsets required by DConv can be obtained from the previous feature maps by other convolutional layers according to

the target task. The channel dimension *2N* corresponds to *N* 2D offsets. *N* is the number of sampling locations in a convolution kernel. A description of the process is shown below.

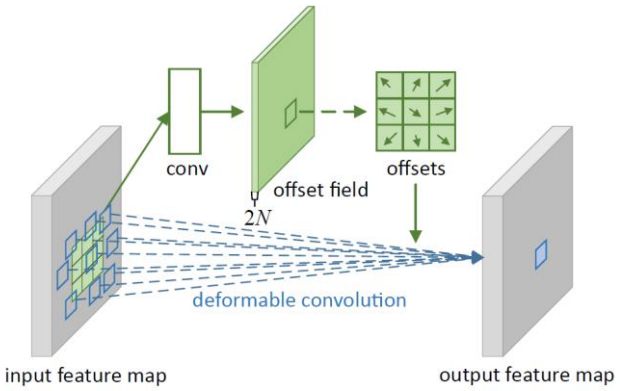


Figure 2.3 Illustration of 3 × 3 deformable convolution [26].

For each location *p* on the output feature map *Y*, a normal convolution process can be expressed as:

$$Y(p) = \sum_{k=1}^{N} \omega_k \cdot X(p + p_k) \qquad (2.1)$$

where X is the input feature map, p_k represents the sampling grid with N sampling locations and ω_k denotes the weights for each location. For example, $N = 9$ and $p_k \in (-1,-1),(-1,0),\dots,(1,1)$ defines a 3×3 convolutional kernel. In the deformable convolution, predicted offsets is added to the sampling grid making deformable kernels spatially-variant. The operation of deformable convolution is as follows:

$$Y(p) = \sum_{k=1}^{N} \omega_k \cdot X(p + p_k + \Delta p_k) \qquad (2.2)$$

where Δp_k is the learnable offset for k-th location. The convolution will be operated on the irregular positions with dynamic weights to achieve adaptive sampling on input features.

Deformable convolution having initially been applied to high-level visual tasks like object detection and semantic segmentation, has shown effectiveness and is now being integrated with video restoration tasks. TDAN [28] first uses DConv in the video super-resolution task to achieve frame alignment. It concatenates the neighboring frame and target frame features, then obtains the sampling offsets through two convolution layers. The offsets are used to replace the traditional optical

flow and align the neighboring frame to the target frame at the feature level, which can be seen in Figure 2.4.

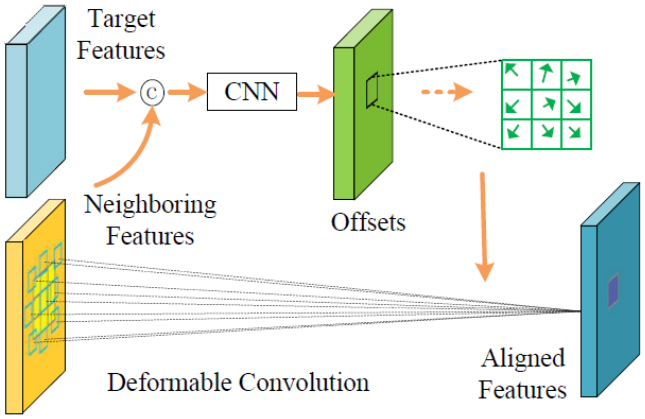


Figure 2.4 Deformable convolution for frame alignment [28].

DConv uses learnable parameters to adjust the kernel to extract lighting and motion features, which can capture complex motion and illumination variations better. Despite the benefits, DConv has limitations, such as high computational complexity and challenging convergence conditions.

2.2 Video Resolution Enhancement Methods

2.2.1 Video Frame Interpolation

Video frame interpolation (VFI) has emerged as a critical technique in video processing that generates intermediate frames between consecutive original frames, resulting in smoother motion and enhanced visual quality. This technology serves multiple practical applications, including slow-motion video creation, frame rate upconversion, and recovery of dropped frames in video transmission systems. The field has seen significant advancements with the rise of convolutional neural networks (CNNs), with current approaches generally categorized into three main methodological frameworks: flow-based, kernel-based, and deformable convolution-based techniques.

Flow-based methods [29-34] leverage optical flow estimation to capture pixel-level motion trajectories between adjacent frames. These approaches first compute dense motion fields that represent how pixels move from one frame to another, then use these motion vectors to guide the synthesis of intermediate frames through warping operations. Early flow-based methods often suffered from occlusion issues and handling of complex non-linear motions, but recent advancements have

introduced sophisticated mechanisms to address these limitations. For instance, some approaches incorporate bidirectional flow estimation, occlusion awareness, and context features to enhance interpolation quality in challenging scenarios involving fast motion or occlusions. Niklaus and Liu [31] propose a context-aware synthesis approach that warps both input frames and their pixel-wise contextual information. As illustrated in Figure 2.5, their method feeds these warped inputs to a frame synthesis neural network, which handles occlusion and large motion better than conventional flow-guided blending methods.

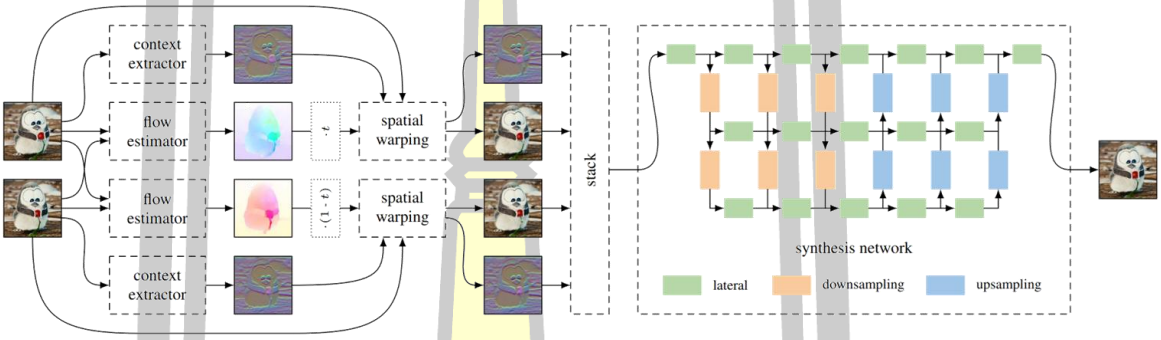


Figure 2.5 The architecture of context-aware synthesis network [31].

Kernel-based methods [35-38] accomplish the interpolation task through adaptive filtering operations. These techniques learn spatially-varying convolution kernels or weighted coefficients that are applied to the input frames to synthesize intermediate content. The key innovation in these approaches is their ability to adaptively determine the contribution of each input pixel to the output frame based on learned contextual information. Advanced kernel-based methods often incorporate multi-scale architectures to handle varying motion magnitudes, enabling them to generate high-quality intermediate frames. Niklaus et al. [36] proposed an adaptive separable convolution approach that estimates pairs of 1D kernels for each output pixel, which can be seen in Figure 2.6. This reduces memory requirements and enables full-frame interpolation with perceptual loss, resulting in visually pleasing results.

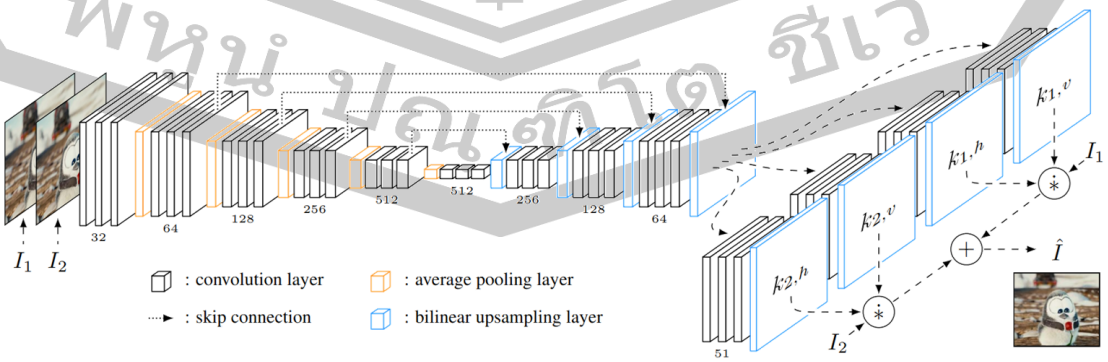


Figure 2.6 The architecture of adaptive separable convolution network [36].

Recently, deformable convolution-based methods [39-43] have gained prominence due to their flexibility in handling complex motion patterns. Unlike standard convolutions with fixed geometric structures, deformable convolutions introduce learnable offsets that enable the network to sample features from arbitrary spatial locations. This enhanced sampling capability allows these methods to better adapt to irregular and complex motion patterns that are common in real-world videos. State-of-the-art deformable convolution-based approaches often combine this flexible spatial sampling with feature pyramids and refinement modules to progressively capture both global and local motion cues, resulting in more temporally coherent and visually pleasing interpolation results. Cheng and Chen [40] proposed DSepConv, which combines deformable convolution with separable kernels. As illustrated in Figure 2.7, their method adaptively estimates kernels, offsets, and masks to capture information from relevant pixels, enabling efficient handling of large motion with small kernel sizes.

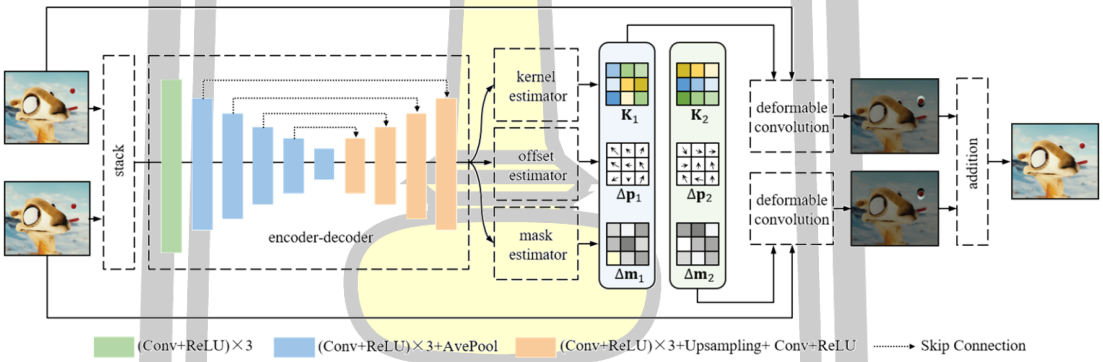


Figure 2.7 The architecture of DSepConv network [40].

Each of these methodological categories presents distinct advantages and limitations, driving ongoing research efforts to develop hybrid approaches that leverage the strengths of multiple paradigms. The continuous evolution of VFI techniques reflects the field's trajectory toward achieving more realistic, artifact-free intermediate frames across diverse video content and motion complexities.

2.2.2 Video Super-Resolution

Video super-resolution (VSR) represents a fundamental challenge in computational video enhancement, aiming to reconstruct high-resolution video content from low-resolution inputs. The core challenge in VSR lies in effectively leveraging temporal correlations between consecutive frames to recover fine details while maintaining temporal consistency. Current deep learning-based VSR methodologies can be categorized into two primary paradigms: alignment-based

methods and non-alignment-based methods, each offering distinct approaches to exploiting inter-frame information.

Alignment-based methods focus on explicit motion modeling between frames, first estimating and compensating for motion before aggregating features across multiple frames. These approaches can be further subdivided based on their alignment mechanisms. Optical flow-based methods [2, 3, 44-47] employ dense motion estimation to establish pixel-level correspondences between adjacent frames and the reference frame. These correspondences guide the warping operations that align features temporally before fusion. Early approaches suffered from inaccuracies in flow estimation, particularly in regions with occlusions or complex motions, but recent advances have incorporated sophisticated flow refinement mechanisms and adaptive feature fusion strategies to mitigate these limitations. Deformable convolution-based methods [1, 28] offer an alternative alignment approach that operates directly at the feature level. By learning spatially-adaptive sampling patterns, these methods can flexibly capture correspondences without explicit motion estimation, enabling more robust handling of complex motions. The landmark works by Chan et al., BasicVSR [3] and its successor BasicVSR++ [48], represent significant milestones in alignment-based VSR, achieving remarkable performance through the integration of bidirectional propagation frameworks with flow-guided deformable alignment modules. The structure of the flow-guided deformable alignment module can be seen in Figure 2.8. It can effectively combine the strengths of both optical flow and deformable convolutions, demonstrating superior capability in handling diverse motion patterns while maintaining computational efficiency.

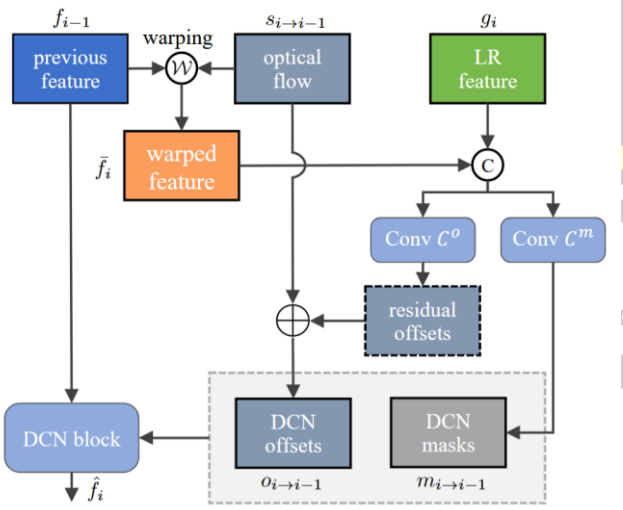


Figure 2.8 Flow-guided deformable alignment [48].

Non-alignment methods take a fundamentally different approach by implicitly modeling temporal correlations without explicit motion compensation. Three-dimensional convolutional networks [49-51] process spatial and temporal dimensions simultaneously through 3D convolutional kernels, enabling the direct extraction of spatio-temporal features. While conceptually straightforward, these approaches often require substantial computational resources and may struggle with large motion displacements. Recurrent network architectures [52-54] offer an alternative non-alignment strategy that progressively accumulates temporal information through recursive processing of frame sequences, which can be seen in Figure 2.9. These methods benefit from parameter efficiency and the ability to capture long-range temporal dependencies, though they may face challenges in propagating information over extended sequences.

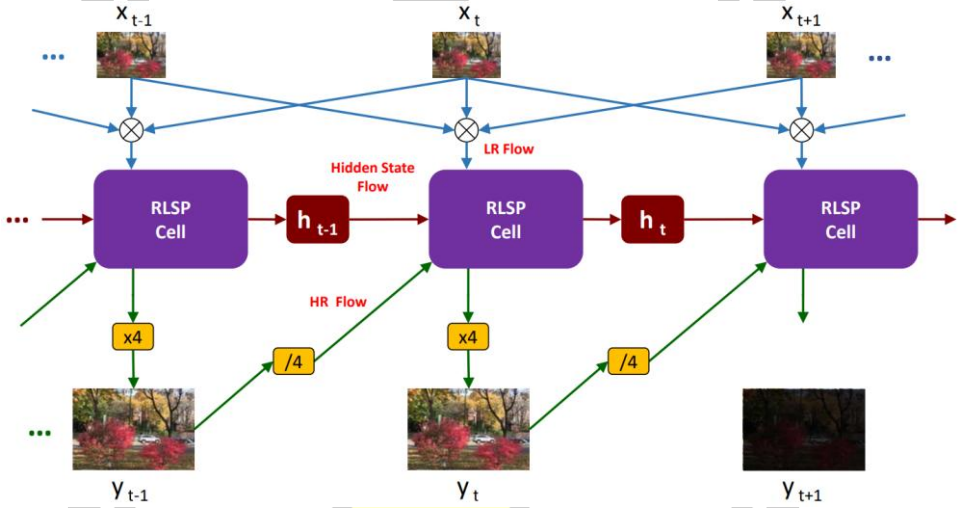


Figure 2.9 Recurrent latent space propagation algorithm [54].

The emergence of transformer-based architectures has catalyzed a paradigm shift in VSR research. Following the groundbreaking introduction of VSRT [25], which demonstrated the potential of self-attention mechanisms for modeling long-range dependencies in video sequences, transformer-based approaches have rapidly evolved. The structure of the spatial-temporal convolutional self-attention used in VSRT can be seen in Figure 2.10. Later, VRT [14] further expanded transformer capabilities beyond super-resolution to simultaneously address multiple video restoration tasks including deblurring and denoising, showcasing the versatility of attention-based frameworks. The subsequent integration of transformers with recurrent architectures [9] represents a significant advancement, effectively combining the global modeling capacity of self-attention with the temporal coherence of recurrent processing. These hybrid approaches have achieved state-of-the-art

performance while maintaining reasonable computational demands through innovative attention mechanisms and architectural optimizations.

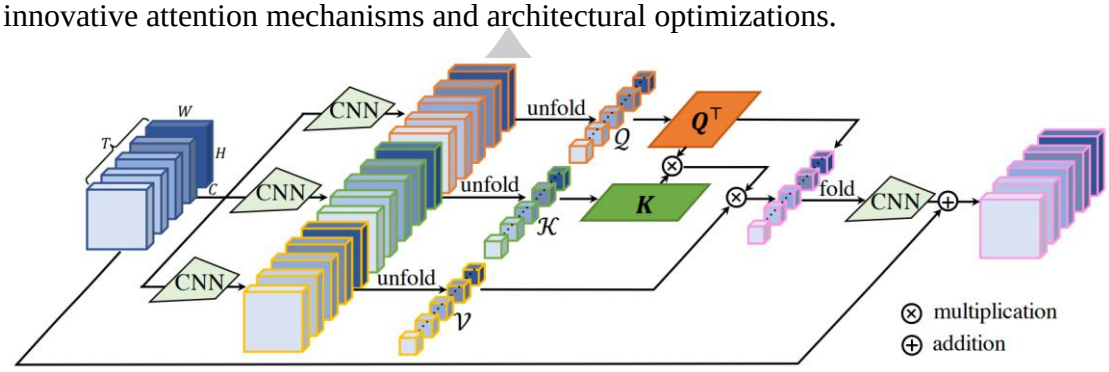


Figure 2.10 Illustration of the spatial-temporal convolutional self-attention [25].

The continuous evolution of VSR methodologies reflects the field's progression toward more sophisticated understanding of spatio-temporal relationships in video sequences. Recent research trends indicate growing interest in developing unified frameworks that can adaptively leverage the strengths of different paradigms, as well as exploring the integration of physical models and learning-based approaches to enhance reconstruction quality while ensuring temporal consistency across diverse video content.

2.2.3 Space-Time Video Super-Resolution

Space-time video super-resolution (STVSR) represents a challenging yet highly impactful domain in video enhancement that simultaneously addresses two critical dimensions of video quality: spatial resolution and temporal frame rate. This dual-objective task aims to transform low-resolution, low-frame-rate video sequences into high-resolution, smooth-motion outputs that deliver significantly enhanced visual experiences. The evolution of STVSR methodologies reflects the field's progression from disjointed processing pipelines to sophisticated integrated frameworks that jointly optimize spatial and temporal enhancements.

Early approaches to STVSR adopted a sequential, multi-stage processing paradigm that treated temporal and spatial enhancement as independent problems. These methods typically began with temporal frame interpolation to generate intermediate frames, followed by spatial super-resolution applied individually to each frame in the sequence. While conceptually straightforward, this cascaded processing suffered from several fundamental limitations [5]. Error propagation became particularly problematic, as artifacts introduced during the interpolation stage were subsequently amplified during the super-resolution phase. Additionally, the independent processing of frames led to temporal inconsistencies, with fluctuations in

quality and appearance across consecutive frames, thereby compromising the perceptual fluidity of the reconstructed videos. These limitations underscored the inherent relationship of the spatial and temporal dimensions in video content and motivated the development of more integrated approaches.

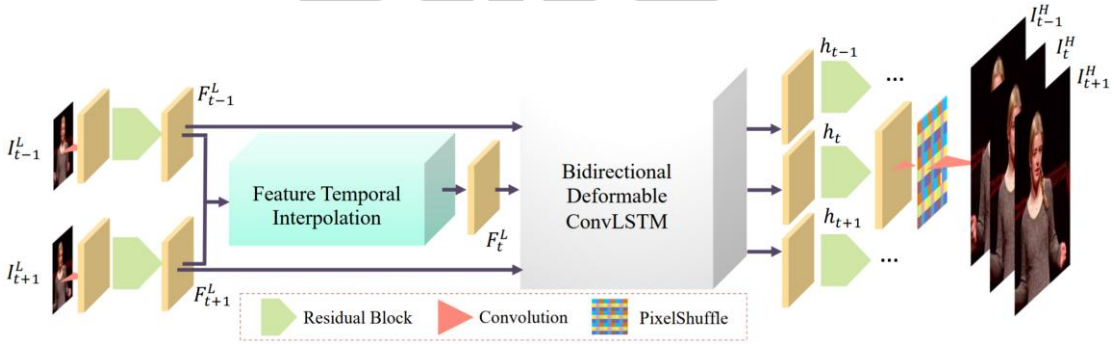


Figure 2.11 The architecture of one-stage STVSR framework [55].

The evolution toward unified STVSR frameworks has been largely facilitated by advances in deep learning architectures. Xiang et al. [55] made a significant contribution with their pioneering one-stage framework Zooming Slow-Mo as illustrated in Figure 2.11. This approach introduced a novel architecture that first performs feature-level frame interpolation using deformable convolution operations to handle complex motion patterns between frames. The intermediate frame features are then processed through a bidirectional deformable ConvLSTM [56] network that effectively captures global temporal correlations across the sequence. This temporal modeling component enables the network to maintain consistency while extracting rich contextual information from neighboring frames. The temporal features are subsequently aggregated and fed into a spatial super-resolution module that reconstructs the final high-resolution frames. By jointly optimizing these components, Zooming Slow-Mo demonstrated superior performance compared to sequential approaches, particularly in preserving temporal coherence and fine spatial details.

Building on this integrated framework concept, Cao et al. [57] introduced an innovative approach that formulates STVSR as an alternating optimization problem with clear mathematical interpretability. Their method uniquely decomposes the complex STVSR task into interrelated sub-problems of deblurring, temporal interpolation, and spatial upscaling, while maintaining tight coupling between these components through an alternating optimization scheme. This formulation allows the framework to address multiple degradation factors simultaneously and adaptively, leading to more robust performance across diverse video content. The interpretable nature of their approach also provides insights into the interaction between spatial and

temporal enhancement processes, facilitating further theoretical developments in the field.

A groundbreaking advancement came with Chen et al.'s [58] introduction of VideoINR, the first continuous space-time super-resolution framework leveraging implicit neural representations. Unlike previous methods that generate outputs at fixed spatial resolutions and frame rates, VideoINR employs a coordinate-based neural network that learns a continuous representation of the video content in both space and time. This implicit representation enables the synthesis of video frames at arbitrary spatial resolutions and temporal positions, providing unprecedented flexibility in video enhancement applications. The continuous nature of the representation also naturally enforces temporal consistency, as frames at any time point are generated from the same underlying function. VideoINR demonstrated remarkable capability in preserving complex motions and fine details while allowing users to freely adjust output specifications according to their requirements or computational constraints.

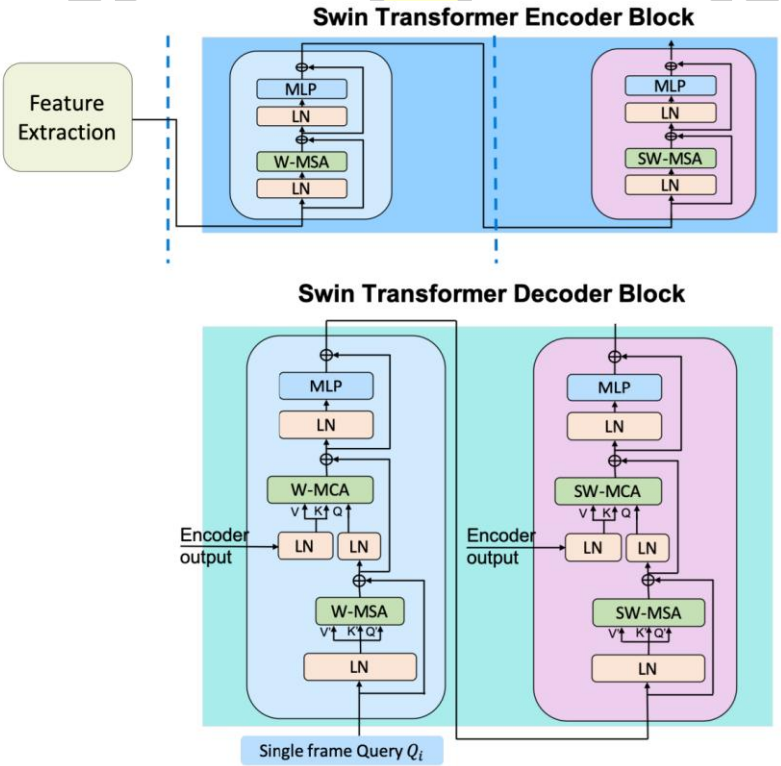


Figure 2.12 The basic swin transformer encoder block and decoder block [59].

The emergence of transformer architectures has further revolutionized the STVSR landscape. Geng et al. [59] proposed RSTT, a transformer-based framework specifically designed to balance computational efficiency with reconstruction quality. As illustrated in Figure 2.12, RSTT employs a single unified transformer architecture that builds reusable dictionaries in encoders based on input frames, which are then

queried in decoders to synthesize high-resolution, high-frame-rate outputs. The proposed approach achieves comparable performance to state-of-the-art methods while being significantly smaller and faster, enabling real-time processing.

The ongoing evolution of STVSR approaches reflects the field's trajectory toward more holistic understanding of video content, where spatial details and temporal dynamics are treated as inherently interconnected aspects rather than separate enhancement targets. Recent research continues to explore more sophisticated architectural designs, novel representations, and efficient computational strategies to push the boundaries of what's possible in joint spatial-temporal video enhancement. The advancements have promising implications for applications ranging from video streaming and archival content restoration to computational cinematography and immersive media experiences.

2.3 Video Deblurring Methods

2.3.1 Single-Image Deblurring

Image deblurring represents a fundamental challenge in computational photography and computer vision, aiming to recover sharp, detailed images from their blurred counterparts. The evolution of this field reflects a notable trajectory from classical optimization-based approaches to sophisticated deep learning architectures, each milestone addressing increasingly complex blur scenarios with enhanced performance.

Traditional image deblurring methodologies predominantly relied on image prior-based formulations [60-62], which approached the task as an optimization problem within a Bayesian framework. These methods typically operated under the assumption of spatially uniform blur kernels, employing carefully designed natural image priors such as sparse gradients, heavy-tailed distributions, or low-rank constraints to regularize the ill-posed inverse problem. While these approaches demonstrated promising results for images affected by camera shake or defocus blur under controlled conditions, they encountered significant limitations when confronted with real-world scenarios involving spatially varying blur patterns resulting from object motion, depth variations, or complex camera movements. The computational intensity of these optimization-based methods also posed practical challenges for real-time applications, often requiring minutes or even hours to process a single image. To address the limitations of uniform blur assumptions, researchers developed more sophisticated approaches incorporating spatial variation in blur modeling. Methods such as [63, 64] introduced segmentation-based strategies that partitioned images into

regions of approximately uniform blur, estimating separate kernels for each segment before integrating the results. This approach improved handling of scene depth variations but often created artifacts at region boundaries. Alternative approaches [65, 66] pursued pixel-level precision by modeling motion fields across the image, effectively estimating a unique blur kernel for each pixel. These methods offered greater flexibility in handling complex motion patterns but increased the parameter space substantially, making optimization more challenging and sensitive to initialization conditions.

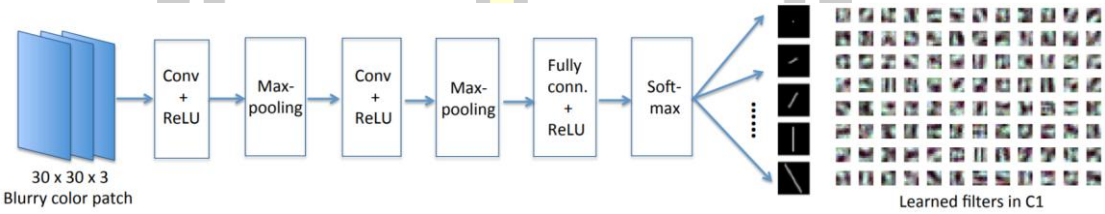


Figure 2.13 Structure of CNN for motion kernels prediction [67].

The rise of deep learning fundamentally transformed the landscape of image deblurring research, enabling data-driven approaches that could implicitly learn complex blur patterns from examples rather than relying on explicit mathematical modeling. Sun et al. [67] pioneered this transition by employing convolutional neural networks to predict non-uniform motion blur kernels at the patch level, which can be seen in Figure 2.13. Their approach combined the predictive power of CNNs with traditional deconvolution techniques, demonstrating that neural networks could effectively estimate spatially varying blur patterns. However, this method still relied on explicit kernel estimation and subsequent deconvolution, inheriting some limitations of classical approaches.

The development of large-scale datasets specifically designed for single image deblurring catalyzed further advancements in end-to-end learning approaches. Several influential works [68-70] demonstrated that CNNs could directly map blurry inputs to sharp outputs without the intermediate step of kernel estimation. These direct methods bypassed the challenges associated with explicit blur modeling and deconvolution, offering more robust performance across diverse blur conditions while significantly reducing computational requirements. Nah et al. [71] made a significant contribution with their multi-scale CNN architecture that progressively refined deblurring results across different resolution scales. As illustrated in Figure 2.14, by processing the input image at multiple resolutions and propagating information from coarse to fine scales, their approach effectively handled a wide spectrum of blur intensities while maintaining computational efficiency. This multi-scale strategy proved particularly effective for severe blur conditions where significant structural information is lost at

the original resolution but may still be partially preserved at lower resolutions. Building upon this multi-scale concept, Tao et al. [72] introduced SRN, which ingeniously integrated recurrent neural network structures within a coarse-to-fine framework. Unlike previous multi-scale approaches that used independent network parameters at each scale, SRN employed shared weights across scales through recursive learning, substantially reducing the parameter count while improving generalization. This architecture enabled effective information propagation between scales, allowing the network to leverage contextual information from coarser resolutions to guide fine-scale restoration, particularly beneficial for regions with large displacement blur.

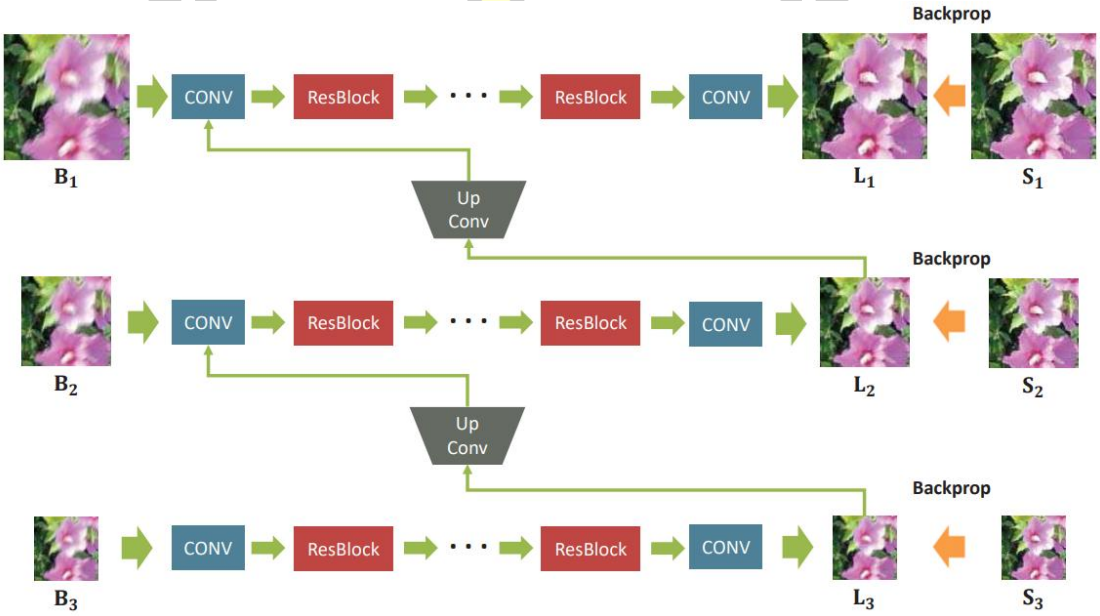


Figure 2.14 The architecture of multi-scale network [71].

The emergence of Transformer architectures in computer vision opened new avenues for image deblurring research by introducing powerful mechanisms for modeling long-range dependencies. Chen et al. [73] pioneered this direction with their image processing transformer, adapting the vision transformer framework to address various image restoration tasks including deblurring. The self-attention mechanisms in transformers enabled the network to capture relationships between distant image regions that might share similar blur patterns or structural characteristics, overcoming the limited receptive field constraints of traditional CNNs. This global modeling capability proved particularly beneficial for handling complex scene structures and non-local blur patterns. Zamir et al. further advanced transformer-based approaches with Restormer [74], which introduced efficient attention mechanisms specifically designed for high-resolution image restoration. As shown in Figure 2.15, Restormer employed a channel-attention transformer architecture that operated on feature

dimensions rather than spatial dimensions, dramatically reducing computational complexity while maintaining the ability to model long-range dependencies. This approach was complemented by progressive feature extraction modules that captured multi-scale information through carefully designed convolutional pathways. The resulting architecture achieved state-of-the-art performance across multiple image restoration tasks, including deblurring, while maintaining practical efficiency for real-world applications.

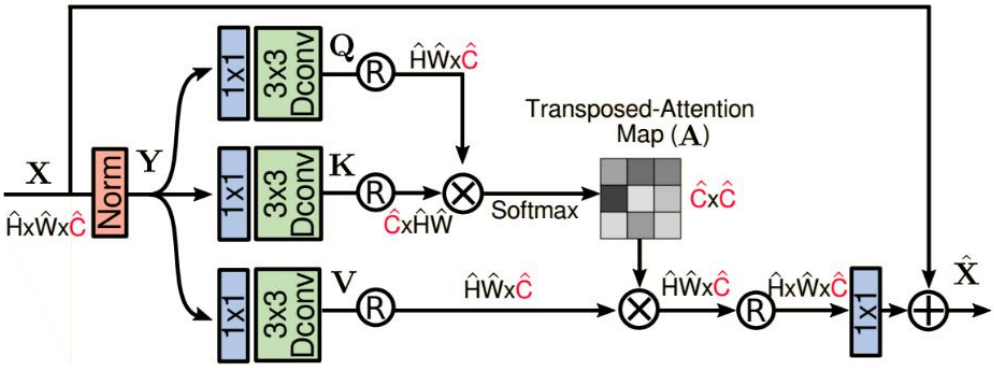


Figure 2.15 Multi-dconv head transposed attention [74].

The ongoing evolution in image deblurring research reflects a continuous pursuit of more effective architectures that can handle increasingly challenging blur scenarios while improving computational efficiency. Recent trends indicate growing interest in hybrid approaches that combine the global modeling capabilities of transformers with the local processing efficiency of CNNs, as well as physics-informed neural networks that incorporate blur formation principles into learning frameworks. These advancements are progressively narrowing the gap between computational deblurring and the human visual system's remarkable ability to interpret blurred visual information, with significant implications for applications ranging from smartphone photography to medical imaging and autonomous vision systems.

2.3.2 Video Deblurring

Video deblurring represents a complex and challenging domain in computational videography that aims to restore clarity and sharpness to degraded video sequences affected by motion blur. Unlike single image deblurring, video deblurring benefits from the temporal redundancy across consecutive frames, offering additional information that can be leveraged to enhance restoration quality. The evolution of this field showcases a remarkable progression from traditional optimization-based methods to sophisticated deep learning architectures that increasingly exploit temporal coherence and long-range dependencies.

In the era of traditional video processing approaches, researchers developed innovative techniques to utilize inter-frame relationships for enhanced blur removal. Matsushita et al. [75] pioneered an influential framework that exploited the observation that not all frames in a video sequence are equally affected by motion blur. Their method identified relatively sharper patches across neighboring frames and intelligently transferred this information to restore blurry regions in the target frame through motion compensation. This approach demonstrated that temporal information could effectively complement spatial processing to overcome the fundamental limitations of single-frame deblurring. Building upon this concept, Cho et al. [76] introduced a more sophisticated registration-based algorithm that modeled patch-based motion trajectories across multiple frames. By accurately tracking the movement of image patches throughout a sequence, their method could better identify and utilize sharp content from adjacent frames, achieving more consistent restoration across dynamic scenes. Kim et al. [77] further advanced this direction by formulating video deblurring as a joint optimization problem that simultaneously estimated motion and recovered sharp content. Their framework integrated motion estimation directly into the restoration process, enabling more accurate handling of complex dynamic scenes with multiple moving objects and camera motion. While these traditional approaches established important foundational principles for leveraging temporal information, they often struggled with computational efficiency and robustness against severe blur conditions or complex motion patterns.

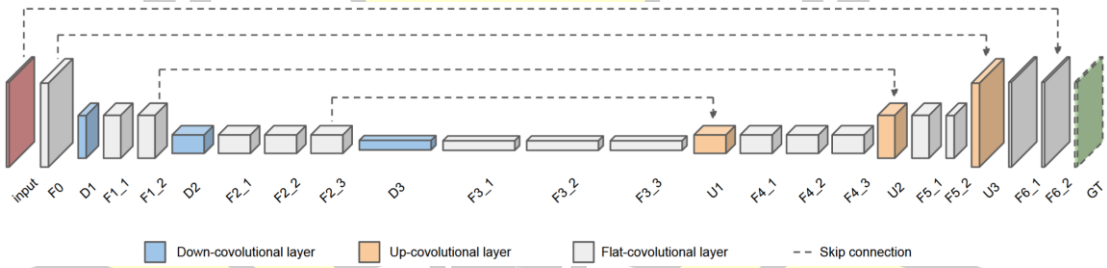


Figure 2.16 The architecture of DeBlurNet [78].

The introduction of deep learning approaches revolutionized video deblurring by offering data-driven solutions that could implicitly learn complex blur patterns and temporal relationships from large-scale datasets. Su et al. [78] made a groundbreaking contribution with the first CNN-based video deblurring framework, as illustrated in Figure 2.16. Their deep encoder-decoder network architecture was specifically designed to process multiple consecutive frames simultaneously, enabling effective exploitation of inter-frame correlations for enhanced restoration. The network learned to align features from adjacent frames and fuse this information to recover details lost in the blurry target frame. This pioneering work demonstrated the significant potential

of deep learning in addressing video-specific challenges that had proven difficult for traditional methods. Zhang et al. [79] further advanced this paradigm by introducing a spatial-temporal CNN architecture that more explicitly modeled the interplay between spatial and temporal domains. Their network incorporated specialized modules for temporal feature extraction and fusion, enabling more effective integration of information across multiple frames while maintaining spatial coherence. This approach significantly improved performance on scenes with complex motion patterns by better preserving temporal consistency in the restored sequence.

The recognition that video deblurring could benefit from modeling longer-term temporal dependencies led to the development of recurrent neural network architectures specifically tailored for this task. The spatial-temporal recurrent network proposed by Kim et al. [80] introduced recurrent connections that enabled information propagation across extended sequences, allowing the network to accumulate and refine features over time. This recurrent processing strategy proved particularly effective for scenes with consistent motion patterns, as the network could progressively enhance its understanding of the underlying sharp content through repeated observations. Building upon this recurrent processing concept, the efficient recurrent neural network introduced by Zhong et al. [81] incorporated LSTM units to more effectively manage the flow of information across time steps, which can be seen in Figure 2.17. The specialized gating mechanisms in LSTM units enabled the network to selectively retain relevant features while filtering out noise, resulting in more stable restoration across long sequences with varying blur conditions. Zhu et al. [82] further refined temporal modeling through a bidirectional feature propagation framework operating across multiple scales. By propagating information both forward and backward in time, and across different resolution levels, their approach captured complementary temporal cues at varying levels of detail, enabling more robust handling of complex motion patterns while preserving fine structural details.

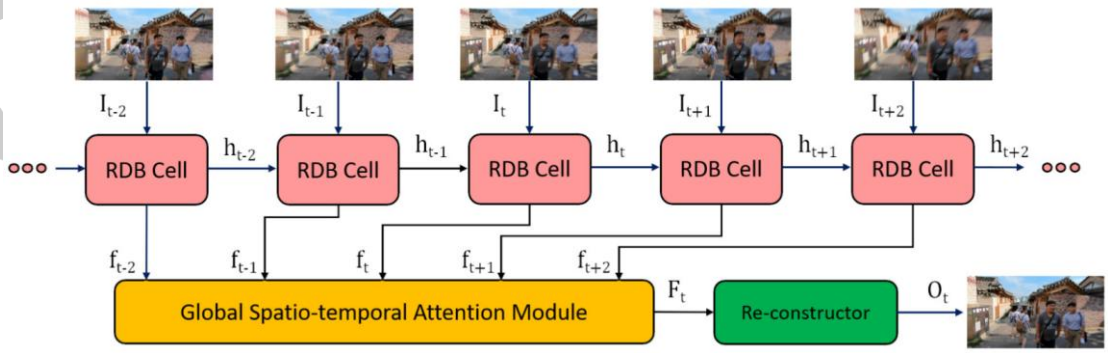


Figure 2.17 The architecture of the efficient recurrent neural network [81].

The integration of attention mechanisms marked a pivotal advancement in video deblurring research, enabling networks to selectively focus on the most relevant spatial and temporal information for restoration. Suin et al. [83] introduced a spatially-attentive patch-hierarchical network that employed sophisticated attention modules to dynamically weight the importance of different spatial regions during feature processing. This selective focus mechanism proved particularly beneficial for scenes with spatially varying blur patterns, as the network could adaptively allocate computational resources according to the complexity and blur severity of different image regions. The attention-guided approach significantly enhanced detail recovery in challenging areas such as textured regions and object boundaries, which had traditionally been problematic for earlier methods. ARVo by Li et al. [84] marked another breakthrough with its innovative spatial correspondence modeling between frames. This framework constructed comprehensive correlation volumes capturing pixel-level relationships across the entire temporal window, as illustrated in Figure 2.18. By first normalizing these correlation volumes to identify strong pixel correspondences, and then selectively aggregating information from neighboring frames based on these correspondence strengths, ARVo effectively transferred sharp details from temporally adjacent frames to restore blurry regions in the reference frame. This sophisticated mechanism for information transfer across frames proved particularly effective for handling complex motion patterns and occlusions, enabling high-quality restoration even under challenging conditions with large displacement blur.

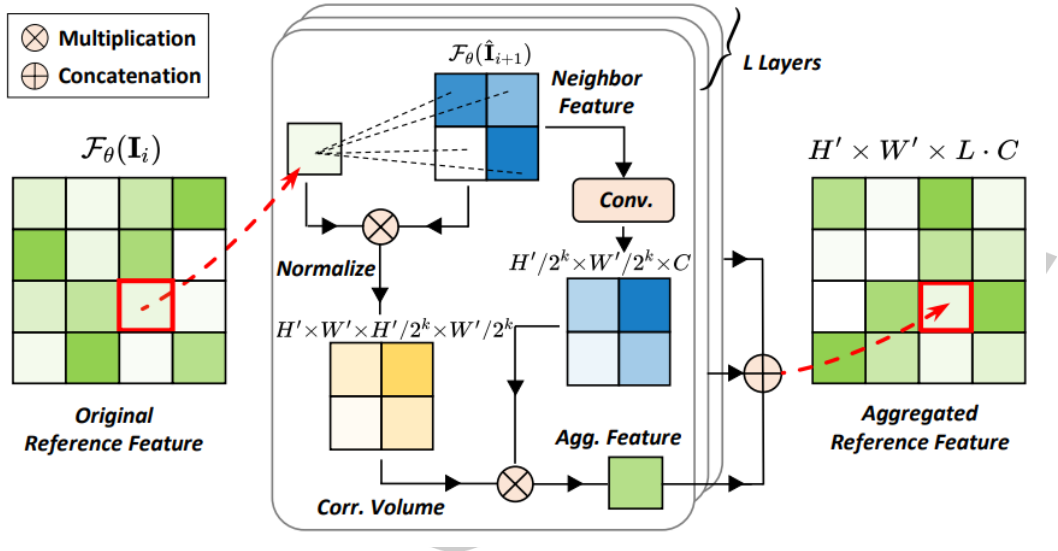


Figure 2.18 Illustration of the correlative aggregation module [84].

The emergence of Transformer architectures in computer vision opened new frontiers for video deblurring, offering powerful mechanisms for modeling long-range

dependencies both spatially and temporally. The video restoration transformer proposed by Liang et al. [14] adapted the transformer paradigm to video restoration tasks including deblurring, introducing specialized attention mechanisms that could efficiently capture relationships between distant pixels across both spatial and temporal dimensions. The self-attention mechanisms in transformers enabled more effective modeling of complex motion patterns and scene structures that extend beyond the limited receptive fields of traditional CNNs. Similarly, the video deblurring transformer by Cao et al. [12] leveraged transformer blocks to establish global dependencies across multiple frames, enabling more coherent restoration of scenes with complex dynamics. These transformer-based approaches demonstrated superior capability in preserving temporal consistency and recovering fine details, particularly in challenging scenarios with rapid motion or significant occlusions. Flow-guided sparse transformer (FGST) by Lin et al. [13] further refined transformer-based video deblurring through an innovative flow-guided sparse attention mechanism, which can be seen in Figure 2.19. By utilizing optical flow to guide the attention process, FGST restricted the attention computation to a sparse set of relevant pixels across frames, substantially reducing computational overhead while maintaining high restoration quality. This approach effectively combined the global modeling capabilities of transformers with the efficiency of local processing, making high-quality video deblurring more accessible for practical applications with computational constraints.

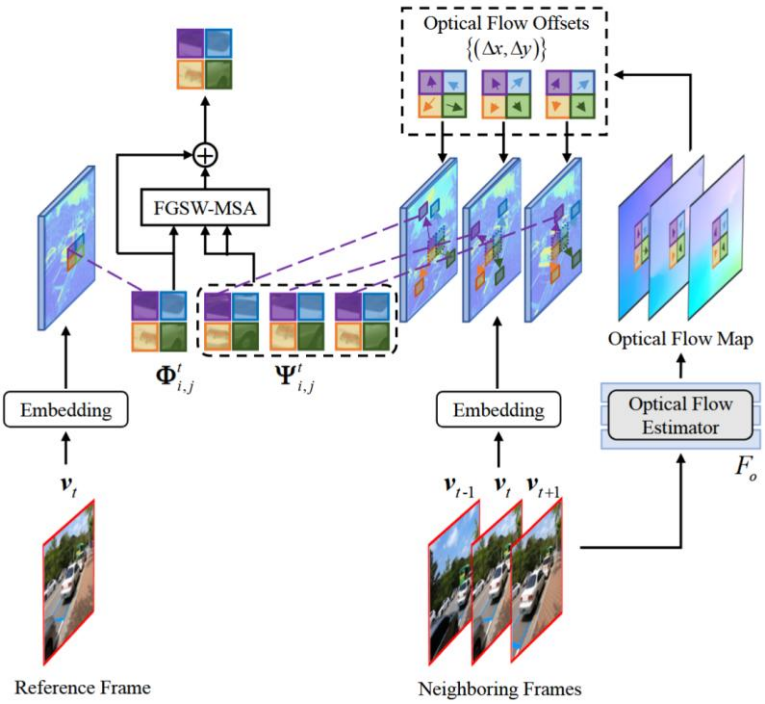


Figure 2.19 The illustration of flow-guided self-attention mechanisms [13].

The ongoing evolution in video deblurring research reflects a continuous refinement of architectural designs to better exploit the unique characteristics of video sequences. Current trends indicate growing interest in hybrid approaches that combine the strengths of different paradigms, such as integrating transformer modules within CNN backbones or incorporating physical blur models to guide the learning process. Additionally, recent work has begun exploring event-based sensing as a complementary modality to standard video frames, leveraging the high temporal resolution of event cameras to better resolve fast motion. These advancements are progressively narrowing the performance gap between computational video deblurring and human perception, with significant implications for applications ranging from consumer videography and film restoration to autonomous driving and surveillance systems that must operate reliably under challenging visual conditions.

2.4 Evaluation Metrics

To measure the quality of generated frames, this work follows the evaluation approach adopted by most state-of-the-art methods. The results will be evaluated quantitatively using peak signal to noise ratio (PSNR) and structural similarity (SSIM). These two widely recognized metrics capture different aspects of image quality assessment. The specific calculation methods of these two indicators are shown below.

2.4.1 Peak Signal to Noise Ratio (PSNR)

Before calculating PSNR, the mean square error (MSE) between the generated image and the reference image should be obtained, which is calculated as follows:

$$MSE = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \| \hat{Y}(i,j) - Y(i,j) \|^2 \qquad (2.3)$$

Here, m and n represent the length and width of the image, while $\hat{Y}(i,j)$ and $Y(i,j)$ represent the pixel values of the generated image and the reference image at point (i,j), respectively. The smaller the value of MSE, the closer the generated image is to the reference image.

The calculation of PSNR is also based on the difference between the generated image and the reference image, and its formula is defined as follows:

$$PSNR = 10 \times \log_{10} \left(\frac{(2^n - 1)^2}{MSE} \right) \qquad (2.4)$$

Where, n is the number of binary bits needed to represent a pixel in the image, namely the bit depth, and $2^n - 1$ is the maximum value of the pixel in the image. The

common image bit depth is 8 bits, the corresponding pixel maximum is 255. Contrary to MSE, the higher the PSNR value, the better the reconstruction quality of the image, in decibels (dB).

2.4.2 Structural Similarity (SSIM)

SSIM is used to measure the structural similarity of two images and is sensitive to local structural changes. SSIM can reflect the similarity degree of images on the whole according to the combination of brightness, contrast, and structure, and its formula is defined as follows:

$$SSIM\ (x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{2.5}$$

Where, *x* and *y* respectively represent two images, *l*, *c*, and *s* respectively represent the comparison of brightness, contrast, and structure of the two images, and the calculation formula is as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2.6}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.7}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{2.8}$$

Where, μ_x and μ_y are the mean values of the two images, σ_x and σ_y are the standard deviations of the two images, σ_{xy} is the covariance of the two images, and C_1, C_2 and C_3 are three constants.

To simplify the form, let $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, then SSIM formula can be expressed as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.9}$$

According to the above formula, the value range of SSIM is [0,1]. The higher the SSIM value, means the higher the similarity between the generated image and the reference image so that the reconstruction quality of the image is better.

2.5 Conclusion

From the above work, we can observe the significant evolution of video restoration research. The field has progressed from traditional methods to advanced deep learning approaches. Different tasks like frame interpolation, super-resolution,

and deblurring have been systematically investigated. Recent developments demonstrate the growing prominence of transformer architectures and attention mechanisms in addressing these challenges. Notably, optical flow estimation and deformable convolution continue to serve as cornerstone techniques for temporal feature modeling. An emerging trend is the development of unified frameworks capable of addressing multiple restoration tasks concurrently.

Chapter 3

Deformable Attention Network for Space-Time Video Super-Resolution

3.1 Introduction

Space-time video super-resolution (STVSR) is a technique that generates high frame rate and high-resolution videos from low frame rate and low-resolution videos. It can improve the visual quality and smoothness of videos, and provide a better viewing experience for people. With the popularity of high-resolution and high refresh rate display devices, this technique has attracted great attention due to its popular applications, such as video compression, video streaming, video analysis, etc.

STVSR is a complicated inverse problem because the given video sequence contains not only spatial blur and noise but also temporal motion blur and frame loss. To solve these problems, it is necessary to consider the spatial-temporal information in the video sequence, that is, the spatial texture details and the temporal motion continuity. With the development of deep convolutional neural networks, the researchers have achieved great success in video restoration tasks, such as video super-resolution (VSR), video frame interpolation (VFI), video denoising and video inpainting. To achieve STVSR, a simple solution is to treat it as a composite task of VFI and VSR. Then, use the current mainstream methods to perform VFI and VSR on low-resolution and low-frame-rate videos sequentially to increase their frame rate and spatial resolution. However, this simple strategy inevitably has some problems. First, the VFI and VSR models will have redundant image feature extraction processes, increasing the overall computational cost. Second, it fails to exploit the intrinsic connection between temporal interpolation and spatial super-resolution. Therefore, this scheme is not the optimal choice.

To better utilize the correlation between the temporal and spatial dimensions in videos, researchers applied a single network for STVSR. Haris et al. [4] proposed a spatial-temporal aware multi-resolution network that utilizes both low-resolution and generated high-resolution frames to capture multi-scale inter-frame motion for intermediate frame synthesis. This method achieves better results than the two-stage methods, but it consumes expensive computation and memory costs. Xiang et al. [55] first uses deformable convolution [27] for video frame interpolation in a one-stage framework, incorporating ConvLSTM [56] for temporal context learning to achieve

state-of-the-art STVSR performance. More recently, Geng et al. [59] proposed RSTT that employs a novel transformer [85] architecture to effectively capture spatial-temporal dependencies, achieving real-time performance while maintaining high reconstruction quality. However, the interdependence of spatial-temporal information poses two critical challenges in STVSR. First, imprecise intermediate frames significantly impact spatial super-resolution quality. Second, insufficient utilization of temporal information across multiple frames prevents achieving optimal reconstruction results. Thus, ensuring the accuracy of intermediate frame generation while improving the efficiency of temporal information extraction remains a crucial problem to be solved.

In this chapter, we propose a novel deformable attention network (DANet). Specifically, in the interpolation module, we introduce a deformable interpolation block (DIB) that enhances inter-frame motion capture, generating more precise sampling parameters for deformable convolution. This enables the module to handle complex inter-frame motion at different scales, thereby improving the synthesis accuracy of intermediate frames. We further incorporate consistency loss to optimize DIB training and enhance module performance. In the temporal fusion module, we design a temporal feature shuffle block (TFSB) to exploit spatial-temporal information across multiple frame features, enabling the network to learn beneficial complementary information between frames. Furthermore, we employ a motion feature enhancement block (MFEB) based on attention mechanism to emphasize motion-related foreground information in features, working synergistically with TFSB to optimize feature learning efficiency.

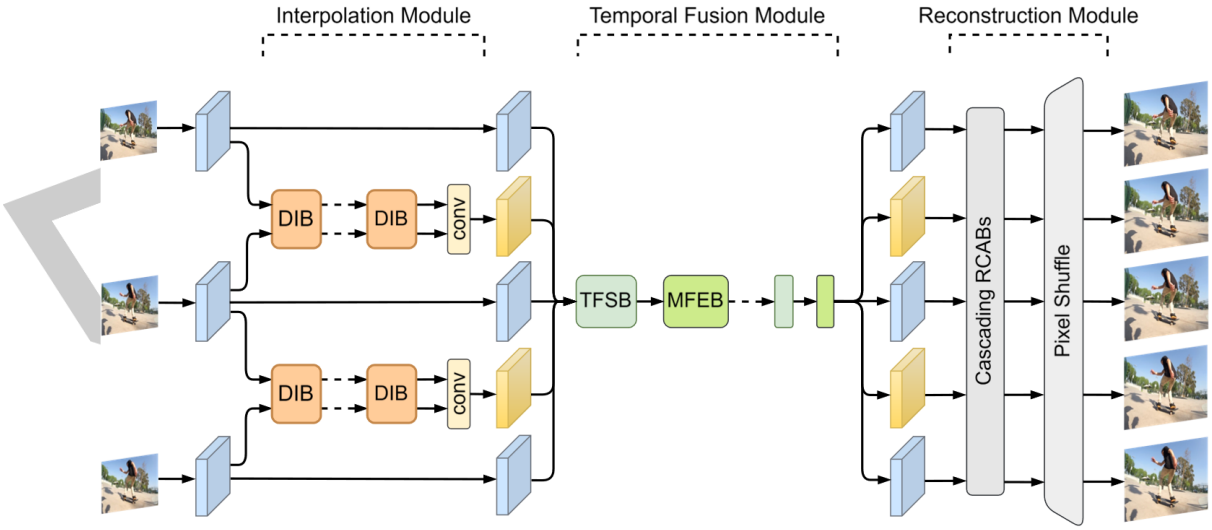


Figure 3.1 The architecture of the proposed deformable attention network (DANet).

3.2 Network Architecture

3.2.1 Network Overview

Given a sequence of low frame rate and low-resolution (LR) video frames $\{I_{2t-1}^{LR}\}_{t=1}^{n+1}$, the proposed network aims to simultaneously improve their spatial and temporal resolution, obtaining a sequence of high frame rate and high-resolution (HR) video frames $\{I_t^{HR}\}_{t=1}^{2n+1}$, where $2n+1$ represents the total number of generated video frames. To efficiently and accurately perform space-time super-resolution on the video in both the spatial and temporal domains, the proposed network architecture is illustrated in Figure 3.1. Considering the computational complexity of the overall network, frame interpolation is first performed on low-resolution inputs, followed by spatial super-resolution. Specifically, the network consists of an interpolation module, a temporal fusion module, and a reconstruction module.

First, shallow features $\{F_{2t-1}^{LR}\}_{t=1}^{n+1}$ are extracted from the input video frames through a convolutional layer followed by several residual blocks. These video frame features are then fed into the interpolation module, where adjacent feature pairs are processed through multiple deformable interpolation blocks (DIB) to generate intermediate frame features $\{F_{2t}^{LR}\}_{t=1}^{n}$, thereby completing the temporal interpolation and yielding high frame rate video features $\{F_t^{LR}\}_{t=1}^{2n+1}$.

Next, to further improve the reconstruction quality of these low-resolution video frame features, the temporal fusion module is employed to facilitate temporal information exchange among the video frame features. This module primarily consists of two key components: the temporal feature shuffle block (TFSB) and the motion feature enhancement block (MFEB). Through the alternating application of these blocks, complementary temporal information can be effectively exchanged among multi-frame features, leading to improved detail reconstruction.

Finally, the reconstruction module, consisting of several residual channel attention blocks (RCAB) [21], extracts high-frequency information from each video frame feature. Through a pixel shuffle operation, this module ultimately generates a sequence of high frame rate, high-resolution video frames.

3.2.2 Interpolation Module

To overcome the limitations caused by optical flow estimation errors, recent approaches have adopted deformable convolution to generate intermediate frames instead of using optical flow. The key challenge in deformable convolution-based frame interpolation lies in generating accurate sampling parameters. Some methods [1,

86] apply deformable convolution sampling on multi-scale features, while others [48, 57] utilize optical flow to guide the generation of sampling parameters for deformable convolution. In this work, we also employ deformable convolution but further enhance its performance by incorporating a hierarchical feature fusion block (HFFB). The HFFB effectively expands the network's receptive field, enabling our proposed deformable interpolation block (DIB) to adaptively handle multi-scale complex motions. Through the cascaded application of multiple DIBs, the interpolation module achieves a coarse-to-fine intermediate frame generation process. Additionally, we introduce a consistency loss during training to constrain the DIBs, further improving the accuracy of generated intermediate frames.

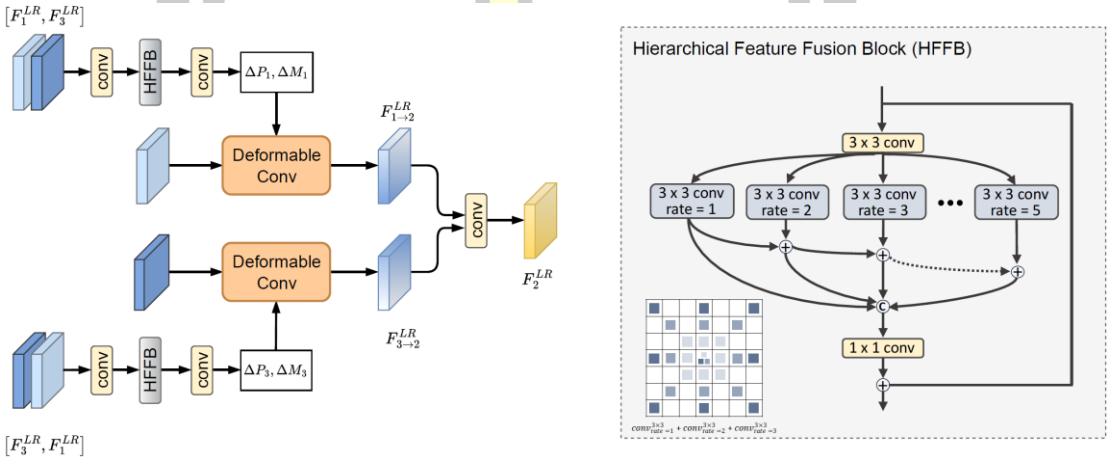


Figure 3.2 The proposed deformable interpolation block (DIB).

The structure of the proposed DIB is illustrated in Figure 3.2. For the input frame features, adjacent frames are fed into the module in pairs to interpolate intermediate frames. Taking F_1^{LR} and F_3^{LR} as examples, they are concatenated in different orders along the channel dimension. To effectively handle inter-frame motions under various scenarios and generate more accurate deformable sampling parameters, we introduce the HFFB. As shown in the figure, HFFB adopts a spatial pyramid structure composed of multiple dilated convolutions with different dilation rates, which efficiently expands the receptive field with relatively low computational cost. In HFFB, feature maps obtained from convolution kernels with different dilation rates are hierarchically summed and then concatenated. The stacking of multiple dilated convolutions not only achieves a larger receptive field but also captures richer feature information. This enables HFFB to more effectively capture pixel-wise motion relationships for generating sampling parameters, facilitating the handling of complex motions at different scales between frames. Subsequently, the corresponding deformable sampling parameters are generated, which can be expressed as follows:

$$\Delta P_1, \Delta M_1 = f([F_1^{LR}, F_3^{LR}]) \tag{3.1}$$

$$\Delta P_3, \Delta M_3 = f([F_3^{LR}, F_1^{LR}]) \quad (3.2)$$

Here, $[\cdot,\cdot]$ denotes feature concatenation, and $f(\cdot)$ represents the sampling parameter generation process, which consists of a 3×3 convolution layer for channel reduction, a HFFB, and a final 3×3 convolution layer that outputs the offset ΔP and modulation factor ΔM. After obtaining the deformable convolution sampling parameters, they are applied to the corresponding input frame features F_1^{LR} and F_3^{LR} to perform forward and backward motion compensation, which can be represented as follows:

$$F_{1\to2}^{LR} = DConv(F_1^{LR}, \Delta P_1, \Delta M_1) \quad (3.3)$$

$$F_{3\to2}^{LR} = DConv(F_3^{LR}, \Delta P_3, \Delta M_3) \quad (3.4)$$

After obtaining two intermediate frame features from the adjacent frame features, these features are fed into the subsequent DIB for further refinement. This cascaded design compensates for potential inaccuracies in motion compensation from a single deformable convolution operation, implementing a coarse-to-fine optimization process. In the last DIB, a 1×1 convolution layer is employed to fuse the two features and output the final intermediate frame feature F_2^{LR}, which can be expressed as follows:

$$F_2^{LR} = Conv_{1\times1}([F_{1\to2}^{LR}, F_{3\to2}^{LR}]) \quad (3.5)$$

Within the DIB, the convolution layers, the HFFB, and the deformable convolution used for differently ordered adjacent frame features share weights. To better train the interpolation module, a consistency loss is introduced as supervision. As shown in Figure 3.3, during the training phase, taking three input frames as an example, after interpolating the intermediate features F_2^{LR} and F_4^{LR}, these two intermediate frame features are further utilized to generate the corresponding $F_3^{LR'}$. The generated features will be closer to the input frame features F_3^{LR} if the interpolation module achieves stronger motion compensation performance. Therefore, an L_1 consistency loss function is employed to constrain the module, enabling the trained DIB to generate more accurate intermediate frames.

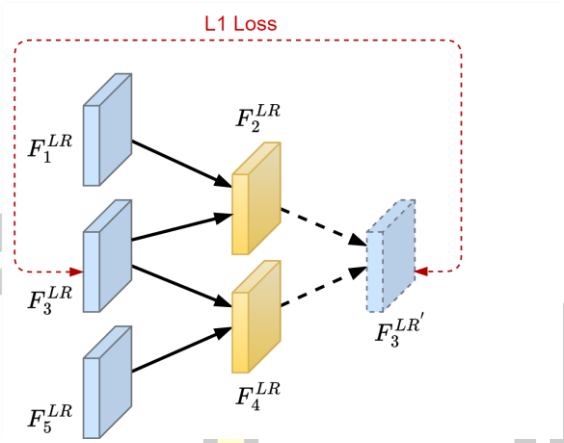
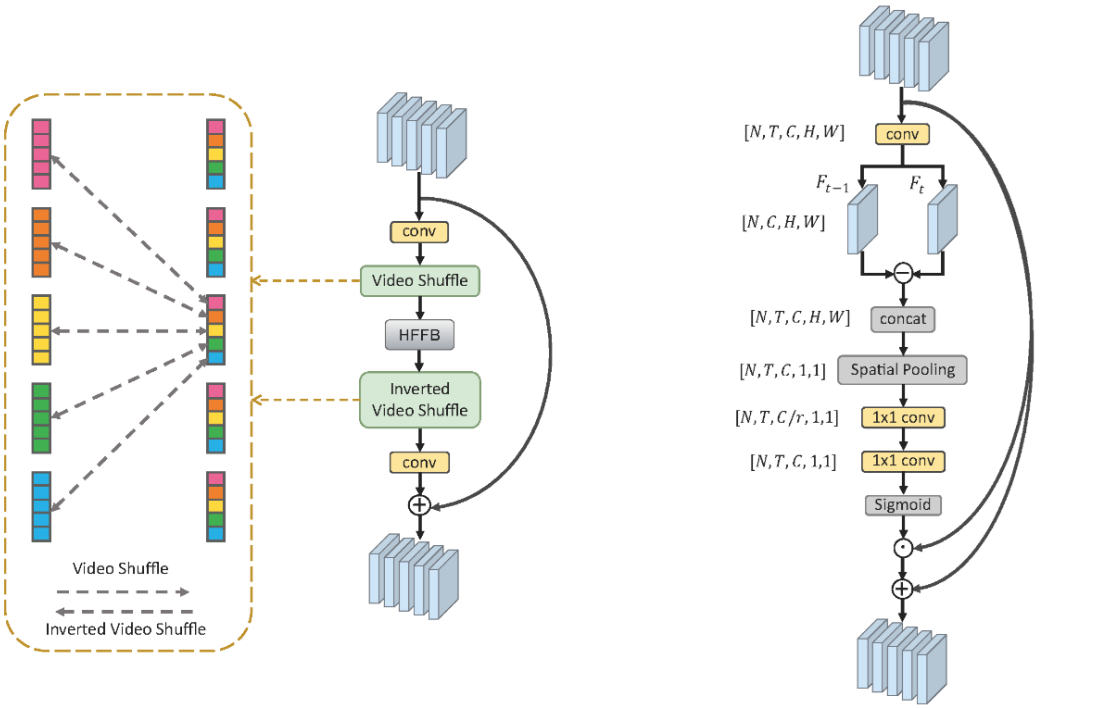


Figure 3.3 Schematic diagram of consistency loss in interpolation module.

3.2.3 Temporal Fusion Module

The temporal information contained within video frame sequences can assist in recovering and restoring missing details. While 3D convolution is commonly used to leverage this temporal information, it significantly increases computational complexity. Therefore, we design a temporal fusion module that efficiently facilitates temporal information exchange among multiple frame features with lower computational costs. As illustrated in Figure 3.4, the module consists of two alternating components: a temporal feature shuffle block (TFSB) based on video shuffle, and a motion feature enhancement block (MFEB) based on attention mechanisms. The MFEB further enhances the inter-frame feature learning efficiency of TFSB, and their synergistic operation enables effective exchange and compensation of missing information among multiple frame features. Next, we will provide a detailed introduction to each of these two blocks.

First is the temporal feature shuffle block (TFSB), which introduces a video shuffle operation. The TFSB introduces a video shuffle operation where each video frame feature is evenly divided into groups corresponding to the number of input frames. For instance, with T input frames of C channels each, the video shuffle operation divides each feature into T groups, each containing $\lfloor C/T \rfloor$ channels. New feature maps incorporating frame features from different time steps are obtained by recombining features from the same group. Temporal information exchange is achieved by applying 2D convolution to these recombined features. Subsequently, an inverted video shuffle operation restores the features to multiple frames, completing one round of temporal information learning.



(a) Temporal Feature Shuffle Block (TFSB) (b) Motion Feature Enhancement Block (MFEB)

Figure 3.4 Schematic diagram of temporal feature shuffle block (TFSB) and motion feature enhancement block (MFEB).

To facilitate the video shuffle operation, a 3×3 convolutional layer is utilized in the TFSB to convert the number of channels to an integer multiple of the input frames, which enables subsequent channel separation. After the video shuffle operation, we employ HFFB again to capture frame-wise feature correlations at different scales, facilitating temporal information extraction and exchange across multiple frames. Subsequently, the multi-frame features obtained from the inverted video shuffle operation are fed into the next 3×3 convolutional layer to restore the original number of channels.

Next is the motion feature enhancement block (MFEB), which is designed based on the observation that different channels in the frame features capture different types of information. Some channels tend to model static information related to the background scene, while others primarily focus on dynamic foreground motion with temporal variations. Considering that the main difference between frames lies in the foreground motion information, which can provide more missing details, the MFEB is designed to enhance these motion features.

The input shape of the MFEB is $[N, T, C, H, W]$. First, these features are passed through a 3×3 convolutional layer, followed by a subtraction operation between each frame feature and its subsequent frame to remove the influence of background

information. For the last frame where no subsequent frame exists, the values are set to zero. The resulting features are concatenated along the temporal dimension as M. To enhance channels that focus on dynamic foreground information, a channel-wise attention mechanism is employed. The specific approach involves applying global average pooling to gather spatial information, which can be represented as:

$$M^s = Pool(M), \quad M^s \in R^{N\times T\times C\times 1\times 1} \tag{3.6}$$

Then, a 1×1 convolutional layer is used to reduce the number of channels by a factor of 16, followed by another 1×1 convolutional layer to restore the original number of channels. After two convolutional layers' non-linear transformations, the Sigmoid function is applied to obtain the weight values for each channel. This process is represented as:

$$S = 2f\left(W_U\delta(W_D M^s)\right) - 1, S \in R^{N\times T\times C\times 1\times 1} \tag{3.7}$$

Here, W_D and W_U represent the two 1×1 convolutions, $\delta(\cdot)$ and $f(\cdot)$ denote the ReLU and Sigmoid functions, respectively. The obtained channel weights S range from $[-1,1]$. Next, the input features are multiplied by their corresponding weights, and the result is added to the input features in a residual fashion, resulting in the final output of the MFEB. This process can be represented as:

$$F^o = F + F \odot S, \quad F^o \in R^{N\times T\times C\times H\times W} \tag{3.8}$$

Here, F represents the input features, $\odot$ denotes element-wise multiplication, and F^o represents the features after motion excitation and enhancement.

3.3 Dataset and Training Details

3.3.1 Dataset

In order to train highly effective VSR networks, it is essential to have a large video dataset for training purposes. Xue et al. [46] assembled a collection of videos from Vimeo and created the Vimeo-90K VSR dataset, which includes 64612 training samples featuring diverse and intricate real-world motions. Each sample is comprised of seven consecutive frames with a standard resolution of 448×256. The Vimeo-90K dataset serves as our primary training dataset. To create LR images, we utilize the MATLAB "imresize" function to perform a $4 \times$ downscaling of the HR images. This process involves initially blurring the input frames using cubic filters and subsequently downsampling them using bicubic interpolation. During training, we select odd-numbered frames as network input, specifically using four LR frames to generate seven HR frames.

For the performance evaluation of the methods, the Vid4 [87], SPMCS [45], Vimeo-90K-T [46] and REDS [88] datasets are used. The Vid4 dataset is widely utilized but is characterized by limited inter-frame motion and contains artifacts on its ground-truth frames. In contrast, the SPMCS dataset offers superior quality video clips showcasing various motions and scenes. The Vimeo-90K-T dataset is notably larger and features a diverse range of flow magnitudes between frames, making it ideal for thorough evaluation of VSR methods. The REDS dataset consists of high-quality 720p video sequences, providing a comprehensive benchmark for evaluating video super-resolution algorithms under real-world scenarios. During the test phase, only the odd frames from video clips are taken as input to generate a complete sequence of video frames.

3.3.2 Training Details

During the training process, we employ data augmentation to enhance the diversity of our training data. This includes techniques such as horizontal or vertical flipping, 90° rotation, and random cropping of the images. Our batch size is set to 8 to optimize the training process. Additionally, the network takes four LR frames as input and extracts image patches of size 50×50 for comprehensive training. The network's output consists of seven corresponding HR image patches. Our model is trained by Adam optimizer [89] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-4} before 60 epochs and later decreases to half every 20 epochs. All experiments were conducted on two NVIDIA RTX 2080 GPUs using PyTorch. The loss function used for training includes the pixel-wise L_1 loss between the seven generated predicted frames and the ground truth HR frames, as well as the consistency loss mentioned in the previous section.

3.4 Experimental Result and Discussion

3.4.1 Comparison with the State-of-the-art Methods

We compare our DANet with several state-of-the-art methods. The evaluation of the results is performed quantitatively using PSNR and SSIM metrics on the Y channel (luminance channel) in the converted YCbCr color space. The comparison methods can be categorized into two groups: (1) two-stage methods that perform frame interpolation followed by super-resolution, and (2) end-to-end space-time super-resolution methods. For the first group, we combine frame interpolation methods (SuperSloMo [30] and DAIN [37]) with super-resolution approaches (RCAN [21], SwinIR [90], RBPN [91] and BasicVSR++ [48]). For the second group, we compare with Zooming Slow-Mo [55], VideoINR [58] and RSTT [59]. These

methods are compared in terms of quantitative results, visual effects, model parameters, and runtime.

1) Comparison of Quantitative Results

The experimental results of various datasets are shown in Table 3.1 and Table 3.2. For two-stage methods, we observe that when using SuperSloMo as the initial frame interpolation method, the video super-resolution (VSR) method RBPN performs even worse than single-image super-resolution methods RCAN and SwinIR on SPMCS and Vimeo-Slow datasets. We attribute this primarily to the fact that although VSR methods can obtain additional temporal information through multi-frame inputs, the use of imperfectly interpolated frames can lead to error propagation, sometimes resulting in degraded reconstruction quality. This is a crucial consideration when employing VSR methods in two-stage tasks.

Among two-stage methods, the combination of DAIN and BasicVSR++ achieves the best results across all datasets. However, this combination generally underperforms one-stage methods on Vid4, SPMCS, and Vimeo test sets. One-stage methods demonstrate significant advantages, as two-stage methods fail to effectively utilize inherent spatial-temporal correlations between frames and cannot fully exploit complementary information. Notably, on the REDS test set, the DAIN and BasicVSR++ combination outperforms all one-stage methods. We attribute this to the characteristics of REDS video clips, which contain larger inter-frame motions and potentially inconsistent camera motion directions, presenting significant challenges for one-stage methods in effectively processing temporal information. BasicVSR++ demonstrates superior capability in handling large-scale motions due to its flow-guided deformable convolution. Our proposed method achieves only marginally lower PSNR (0.06dB difference) while maintaining the best PSNR across other datasets, demonstrating the effectiveness and strong adaptability of our method across diverse scenarios.

Table 3.1 Quantitative comparison of state-of-the-art methods on Vid4, SPMCS and REDS datasets for 4 ×.

VFI Method	VSR Method	Vid4		SPMCS		REDS	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SuperSloMo	RCAN	23.78	0.6383	28.11	0.7763	26.57	0.7300
SuperSloMo	SwinIR	23.90	0.6438	28.36	0.7826	26.72	0.7334
SuperSloMo	RBPN	24.17	0.6584	27.92	0.7759	26.63	0.7338
SuperSloMo	BasicVSR++	24.29	0.6894	28.18	0.7854	26.75	0.7353
DAIN	RCAN	24.06	0.6759	28.20	0.7935	26.83	0.7446
DAIN	SwinIR	24.19	0.6819	28.45	0.8017	26.99	0.7483
DAIN	RBPN	24.08	0.6859	26.85	0.7618	27.05	0.7534
DAIN	BasicVSR++	24.41	0.7020	28.74	0.8131	**27.31**	0.7610
Zooming Slow-Mo		26.38	0.7977	30.95	0.8731	27.16	0.7635
VideoINR		26.47	0.8024	31.28	0.8793	26.98	0.7579
RSTT		26.43	0.7993	31.15	0.8750	27.02	0.7586
Ours		**26.58**	**0.8057**	**31.50**	**0.8822**	27.25	**0.7639**

Table 3.2 Quantitative comparison of state-of-the-art methods on Vimeo-90K-T dataset for 4 ×.

VFI Method	VSR Method	Vimeo-Slow		Vimeo-Medium		Vimeo-Fast	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SuperSloMo	RCAN	30.80	0.8642	32.51	0.8883	34.51	0.9075
SuperSloMo	SwinIR	30.99	0.8676	32.71	0.8908	34.71	0.9092
SuperSloMo	RBPN	30.62	0.8625	32.88	0.8943	34.89	0.9135
SuperSloMo	BasicVSR++	31.12	0.8684	33.97	0.9004	35.22	0.9149
DAIN	RCAN	32.29	0.8996	33.80	0.9141	35.31	0.9251
DAIN	SwinIR	32.48	0.9004	34.00	0.9162	35.54	0.9266
DAIN	RBPN	33.08	0.9112	34.48	0.9270	35.61	0.9314
DAIN	BasicVSR++	33.24	0.9125	34.75	0.9288	36.03	0.9347
Zooming Slow-Mo		33.38	0.9139	35.42	0.9362	36.81	0.9415
VideoINR		33.40	0.9144	35.49	0.9364	36.72	0.9389
RSTT		33.55	0.9151	35.62	0.9371	36.79	0.9397
Ours		**33.67**	**0.9179**	**35.64**	**0.9376**	**36.98**	**0.9431**

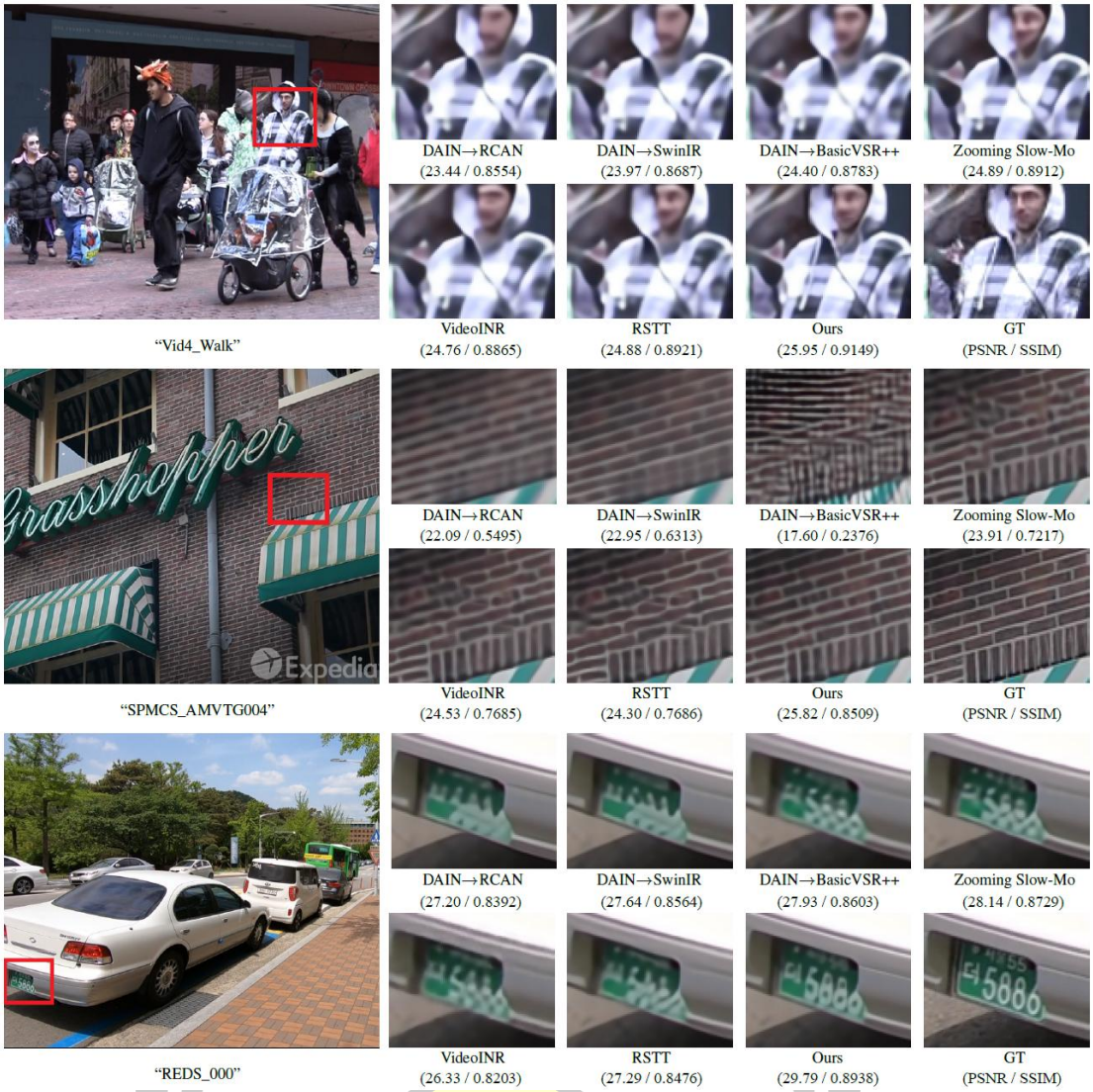


Figure 3.5 Visual results on Vid4, SPMCS and REDS for 4 × scaling factor.

2) Comparison of Qualitative Results

The video frame sequences generated after spatial-temporal super-resolution can be divided into two categories: original video frames that only undergo spatial super-resolution and intermediate frames that undergo both spatial and temporal super-resolution. Generally, the quality of the intermediate frames is inferior to that of the original frames. Low-quality intermediate frames introduce flickering artifacts that have a greater impact on the overall visual perception of the video. Therefore, to better compare the methods, the following visual comparison results are selected from the intermediate frames.

Figure 3.5 demonstrates the visual comparison between our proposed method and other two-stage and one-stage methods on Vid4, SPMCS and REDS datasets. To highlight the differences in detail between the generated video frames, specific

regions are boxed and magnified for comparison in the figures. In the "Vid4_Walk" clip, other methods fail to recover the drawstring beneath the hoodie, while our method successfully reconstructs its contours. In the "SPMCS_ANVTG004" clip, only the proposed method exhibits brick wall textures that are closest to the real images. The results obtained by other methods show varying degrees of texture misalignment and partial blurriness in certain regions, particularly in frames generated by DAIN → BasicVSR++, where severe texture distortion is observed. In the "REDS_000" clip, while some methods can vaguely show the license plate numbers, the clarity is insufficient with blurred edges. Our method provides better recognition of the specific numbers.

Figure 3.6 Visual results on Vimeo-90K-T for 4 × scaling factor.

Figure 3.6 presents the visual comparison of the methods on the Vimeo-90K-T dataset. The first two rows compare the texture restoration capabilities of different methods. Our proposed method generates the sharpest and most accurate venetian blind stripes, and clearer logo lines, demonstrating superior detail preservation. Other methods exhibit varying degrees of texture artifacts. The latter two rows compare text restoration capabilities across methods. The proposed method achieves edge sharpness and clarity closest to the ground truth, making text more legible. Other methods introduce artifacts not present in the original frames to varying degrees, compromising text readability.

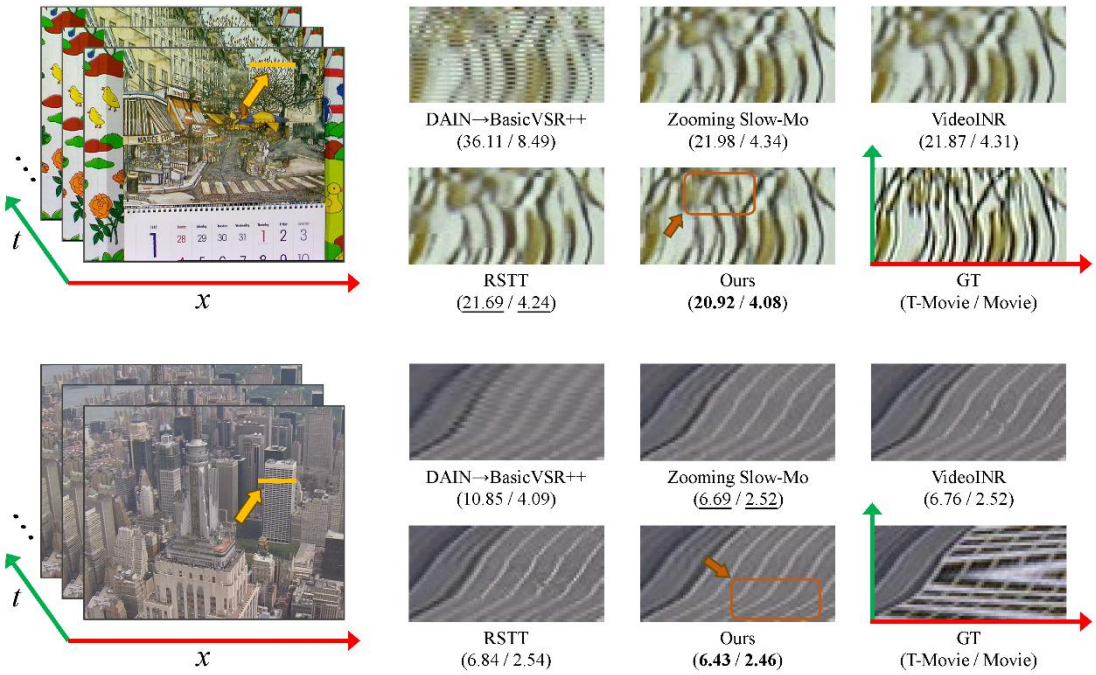


Figure 3.7 Comparison of time domain profiles generated by different methods of video clips.

3) Temporal Consistency Comparison

In addition to comparing the detailed texture of the generated video frames on multiple datasets, Figure 3.7 demonstrates the ability of different methods to produce temporally consistent results. This is achieved by extracting pixels from the same horizontal line across consecutive video frames (indicated by the orange line in the figure) and vertically stacking them to generate temporal profile plots. The results in the figure indicate that the proposed method exhibits good temporal consistency in the generated video frames, with fewer occurrences of flickering artifacts in the images. Notably, the second comparison set displays the reconstruction of building window grids, showing significant differences from the ground truth. These fine grid structures undergo substantial information loss after 4 × downsampling, posing a major challenge for super-resolution tasks. Consequently, existing methods can only partially recover the original window details. Besides visual comparisons, we provide T-Movie and Movie [92] metrics beneath each image. The Movie metric evaluates video quality by jointly considering spatial and temporal aspects. T-Movie serves as the temporal component of the Movie metric, specifically focusing on assessing the continuity and temporal consistency between video frames. Lower values in both metrics indicate better video quality. Our proposed method also achieves the best performance on both metrics.

Table 3.3 Comparison of parameters and runtime for different methods on the SPMCS Dataset.

Method	Parameters (M)	Total Runtime (s)	Runtime (s/frame)	PSNR	SSIM
SuperSloMo → RCAN	19.8 + 16.0	0.47 + 19.87	0.656	28.11	0.7763
SuperSloMo → SwinIR	19.8 + 11.9	0.47 + 434.53	14.032	28.36	0.7826
SuperSloMo → RBPN	19.8 + 12.7	0.47 + 83.17	2.698	27.92	0.7759
SuperSloMo → BasicVSR++	19.8 + 7.3	0.47 + 4.48	0.160	28.18	0.7854
DAIN → RCAN	24.0 + 16.0	4.59 + 19.87	0.789	28.20	0.7935
DAIN → SwinIR	24.0 + 11.9	4.59 + 434.53	14.165	28.45	0.8017
DAIN → RBPN	24.0 + 12.7	4.59 + 83.17	2.831	26.85	0.7618
DAIN → BasicVSR++	24.0 + 7.3	4.59 + 4.48	0.293	28.74	0.8131
Zooming Slow-Mo	11.1	4.37	0.141	30.95	0.8731
VideoINR	11.3	3.96	0.128	<u>31.28</u>	<u>0.8793</u>
RSTT	7.7	4.55	0.147	31.15	0.8750
Ours	8.8	3.81	0.123	**31.50**	**0.8822**

4) Model Size and Running Time Comparison

Table 3.3 presents a comparison of model sizes and runtime for all methods. In the case of two-stage method combinations, the parameter size and runtime are the sum of the corresponding values for the two models. The testing is conducted on the SPMCS dataset, which consists of 32 video clips, with each clip containing 16 video frames of size 240×135. After the spatial-temporal super-resolution process, 31 video frames of size 960×540 are generated. The average time taken for each clip in this process is calculated, resulting in the total runtime for processing one clip. The total runtime is then divided by 31 frames to obtain the average time required to generate each HR video frame.

As observed from the table, two-stage method combinations exhibit significantly larger parameter sizes compared to one-stage methods. Furthermore, two-stage methods require longer processing time due to the additional overhead of feature extraction performed by two separate networks. Compared to other one-stage methods, our proposed method achieves higher quality video frame reconstruction

with shorter execution time, demonstrating its effectiveness and superiority in terms of computational efficiency.

3.4.2 Ablation Study

To further investigate the proposed method, this section conducts ablation experiments by analyzing the effectiveness of the designed modules in the network. The capability of the interpolation module to reconstruct accurate intermediate frame features significantly impacts the subsequent frame reconstruction quality. The temporal fusion module determines whether inter-frame features can effectively acquire complementary temporal information. Therefore, the focus here is on the design and analysis of the ablation experiments for the interpolation module and the temporal fusion module.

1) Effectiveness of Components in the Interpolation Module

The interpolation module aims to synthesize intermediate frame features from adjacent frame features. Since this module utilizes DIB for feature-level inter-frame motion compensation, the impact of the number of DIB used in the module is explored. The results of the ablation experiments are shown in Table 3.4. In the table, "0 DIB" represents not using DIB for motion compensation but instead employing traditional optical flow methods. Specifically, PWC-Net [93] is first used to estimate the forward and backward optical flow between adjacent frames and then the intermediate frame is synthesized. It can be observed that using one DIB achieves better results than the optical flow method, with a PSNR improvement of 0.11 dB, and without the need for an additional optical flow estimation network. This demonstrates that the proposed structure can achieve the intended purpose effectively. When the number of DIB is increased to two, the model's performance is further improved, with a PSNR improvement of 0.15 dB. However, there is no significant improvement in the results when the number of DIB reaches three, indicating that two DIB are sufficient to generate relatively accurate intermediate frames. Therefore, the proposed method selects two DIB in the interpolation module. As mentioned earlier, to better train the interpolation module and generate more accurate intermediate frames, we add consistency loss during the training process to constrain the module and accelerate its convergence. In the ablation experiments, using two DIB and adding consistency loss, the results in the table further improve by 0.12 dB, demonstrating the effectiveness of this loss.

Table 3.4 Ablation study of the interpolation module on the Vimeo-Fast dataset.

0 DIB	1 DIB	2 DIB	3 DIB	Consistent Loss	Params (M)	FLOPs (G)	PSNR	SSIM
✓					7.86	288.68	36.60	0.9359
	✓				8.33	300.65	36.71	0.9385
		✓			8.81	312.61	36.86	0.9410
			✓		9.29	324.58	36.90	0.9422
		✓		✓	8.81	312.61	**36.98**	**0.9431**

2) Effectiveness of Components in the Temporal Fusion Module

The temporal fusion module is responsible for fully exploiting the spatio-temporal complementary information among multi-frame features. It consists of two alternately cascaded blocks: the temporal feature shuffle block (TFSB) and the motion feature enhancement block (MFEB). Ablation experiments were conducted on these two blocks, with results shown in Table 3.5. It can be observed that the network performs poorest when directly feeding interpolated features to the reconstruction module without the temporal fusion module. Adding only MFEB results in minimal performance improvement. This is because MFEB's role is to enhance motion features containing richer temporal information to facilitate inter-frame feature learning by TFSB. Without TFSB, the sole addition of MFEB provides limited benefits to the network. When only TFSB is added, although the network's parameters and computational cost increase notably, the PSNR value improves by 0.26 dB. The subsequent integration of MFEB achieves an additional 0.15 dB improvement with minimal computational overhead, demonstrating its efficiency in performance enhancement.

Table 3.5 Ablation study of the temporal fusion module on the Vimeo-Fast dataset.

TFSB	MFEB	Params (M)	FLOPs (G)	PSNR	SSIM
		5.40	203.42	36.57	0.9351
	✓	5.72	210.58	36.64	0.9369
✓		8.49	305.45	36.83	0.9402
✓	✓	8.81	312.61	**36.98**	**0.9431**

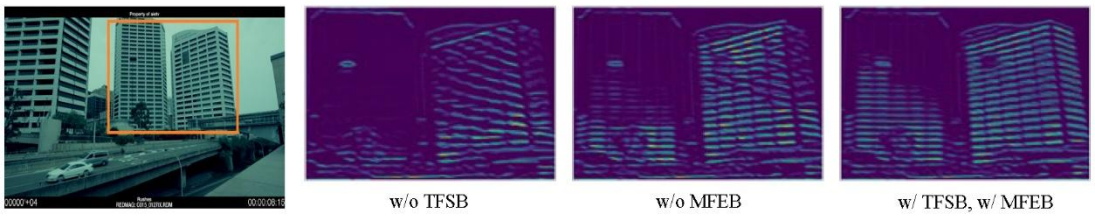


Figure 3.8 Feature maps visualization of ablation on temporal fusion module.

To better demonstrate the effectiveness of both blocks in the temporal fusion module, we visualize the feature maps under different configurations. As shown in Figure 3.8, two buildings move with the camera motion. Without TFSB, the network fails to learn temporal information contained in other frames, resulting in mostly incorrect building structure and texture in the features. Without MFEB, although inter-frame features undergo temporal information exchange and learning, some texture details remain missing or blurred. When TFSB and MFEB are used together, the two blocks work collaboratively to produce more complete building structure edges and texture information. The above analysis demonstrates the effectiveness of both blocks' design in the temporal fusion module, and verifies the importance of their complementarity and synergistic effects.

Table 3.6 Experimental results with a different number of input frames.

Input	Output	PSNR	SSIM
2 frames	3 frames	36.72	0.9387
3 frames	5 frames	36.89	0.9417
4 frames	7 frames	**36.98**	**0.9431**

3) Impact of Frame Count

To investigate the impact of inter-frame information on restoration results, we trained the network with varying numbers of input frames. Since each training sample in the Vimeo-90K dataset contains only 7 consecutive video frames, we conducted comparisons using 2 to 4 input frames. As shown in Table 3.6, the network's performance improves with an increasing number of input frames, as it can leverage more beneficial temporal information for frame reconstruction. However, the performance gain demonstrates diminishing returns when increasing from 3 to 4 input frames, indicating that the benefits of additional temporal information become less significant with more frames. For STVSR task, the last frame of the current input becomes the initial frame of the next input sequence, meaning this frame undergoes reconstruction twice. With more input frames, the computational overhead relative to the entire video clip decreases. Therefore, we choose 4 frames as the network input.

3.4.3 Limitations

Although the proposed method demonstrates good video spatio-temporal super-resolution effects in most scenarios, it still faces some notable limitations, as shown in Figure 3.9. When processing scenes containing regular repetitive textures (such as brick walls) and complex geometric structures (such as building facades), our method struggles to fully recover high-frequency details and precise geometric features, resulting in blurred edges, structural distortion, or over-smoothing in the reconstructed results. Such real-world scenes with complex details present a significant challenge for all spatio-temporal super-resolution methods, as they require algorithms to simultaneously handle rich texture information and precise geometric correspondences. Despite significant advances in current deep learning frameworks, achieving realistic and accurate reconstruction in these scenarios still requires exploration of more effective feature extraction and spatio-temporal fusion strategies, possibly incorporating stronger prior knowledge or more refined motion estimation mechanisms to overcome these inherent difficulties.

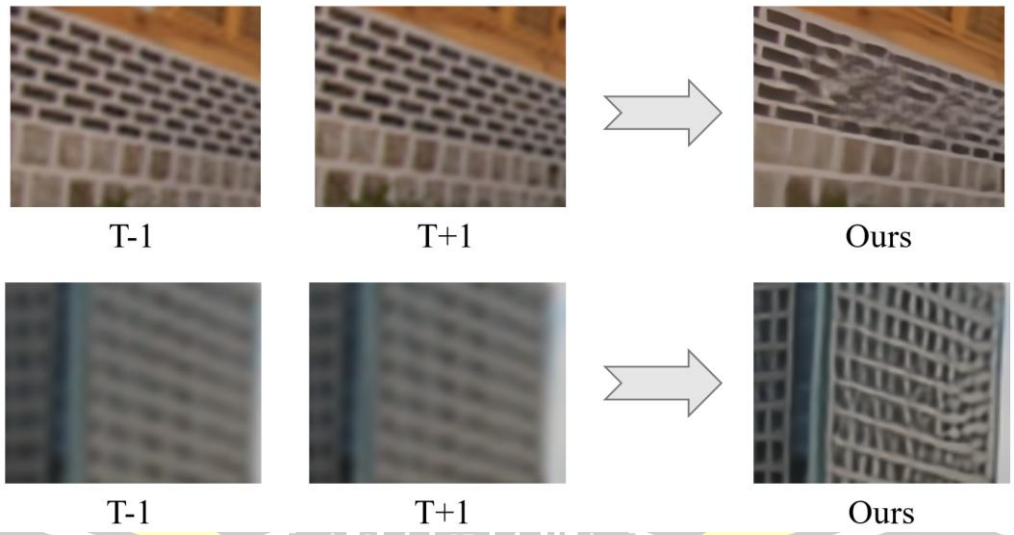


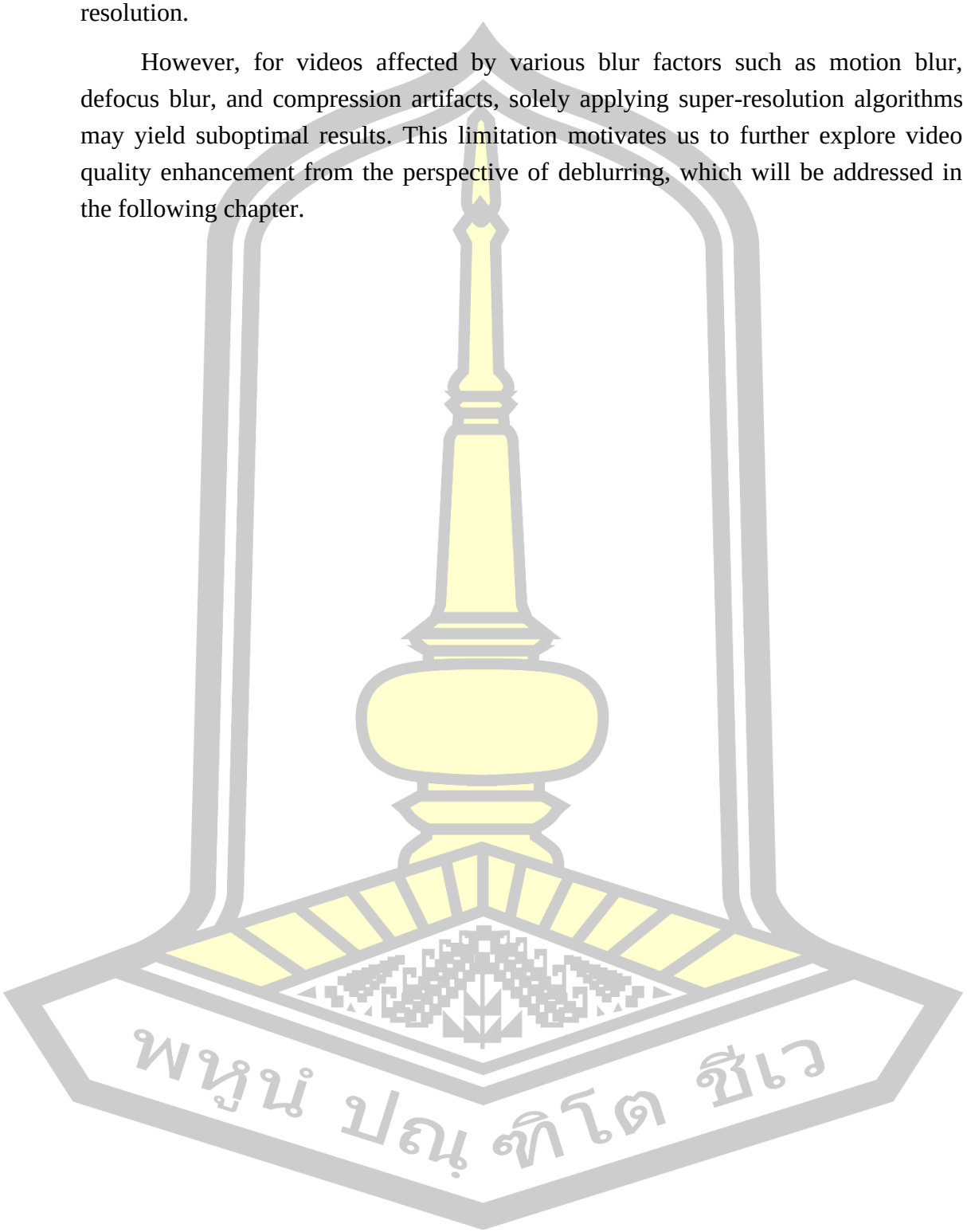
Figure 3.9 The failure cases of the DANet.

3.5 Conclusion

In this chapter, we propose a novel deformable attention network (DANet). In the interpolation module, we design a deformable interpolation block (DIB) to adapt to various complex inter-frame motion scenarios and interpolate accurate intermediate frame features. In the temporal fusion module, we propose a temporal feature shuffle block (TFSB) and a motion feature enhancement block (MFEB) to effectively enhance inter-frame motion features and learn complementary temporal information between frames. The extensive experiments on benchmark datasets definitively prove

the remarkable effectiveness of our DANet in enhancing video space-time super-resolution.

However, for videos affected by various blur factors such as motion blur, defocus blur, and compression artifacts, solely applying super-resolution algorithms may yield suboptimal results. This limitation motivates us to further explore video quality enhancement from the perspective of deblurring, which will be addressed in the following chapter.



Chapter 4

Wavelet-Based Blur-Aware Decoupled Network for Video Deblurring

4.1 Introduction

While the previous chapter focused on video super-resolution, we recognize that real-world videos often suffer from multiple degradation factors simultaneously. Among these factors, blur is a particularly common issue that significantly impacts video quality. The proliferation of mobile smart devices has made video recording an essential part of daily life, yet captured videos frequently experience blur as a result of camera shake and object motion. Therefore, the development and application of effective video deblurring algorithms are highly beneficial not only for improving video quality but also for enhancing the overall viewing experience.

Video deblurring remains a challenging restoration task, with its core difficulty lying in effectively utilizing temporal information from video sequences. Research [94] has shown that appropriate feature alignment strategies are vital for improving deblurring performance. Early methods primarily employed optical flow [6-8] for explicit alignment, but blur-induced inaccuracies in flow estimation propagated errors that compromised reconstruction quality. To address this limitation, researchers developed adaptive feature alignment schemes based on deformable convolution [9-11]. Recently, transformers have emerged as a promising solution [12-14] for video deblurring, demonstrating superior performance in complex scenarios through their powerful self-attention mechanisms and long-range dependency modeling capabilities.

However, video deblurring faces unique challenges compared to other video restoration tasks, like super-resolution. Blur not only results in the loss of high-frequency details but also significantly distorts the fundamental image structure. While existing methods have achieved remarkable progress in detail recovery, they often overlook the equally important aspect of structural information reconstruction. Therefore, effectively addressing the simultaneous management of high-frequency details and low-frequency structural information in video deblurring continues to pose a significant challenge within the field of research.

To address these issues, this chapter proposes a wavelet-based, blur-aware decoupled network (WBDNet) that decouples structure recovery from detail enhancement. Unlike previous approaches [95-97] that merely used wavelet

transforms for multi-scale feature extraction, we design specialized recovery strategies for different frequency bands. Specifically, we introduce a multi-scale progressive fusion (MSPF) module and a blur-aware detail enhancement (BADE) module.

In the low-frequency domain, the MSPF module operates in several steps. It first employs optical flow for coarse alignment. Then, it constructs multi-scale feature pyramids and uses bottom-up progressive feature fusion to reconstruct the main image structure. In the high-frequency domain, the BADE module takes a different approach. It integrates blur-aware attention mechanisms with deformable convolution. This integration enhances features in sharp regions, thereby enabling more refined alignment. As a result, this approach enables the effective extraction and integration of valuable detailed information from multiple frames.

4.2 Network Architecture

4.2.1 Network Overview

As shown in Figure 4.1, given a sequence of blurred video frames $\{I_t^{Blur}\}_{t=1}^{N}$, our goal is to reconstruct the corresponding sharp video frames $\{I_t^{Sharp}\}_{t=1}^{N}$. The proposed network architecture consists of two main branches: a preprocessing branch and a feature reconstruction branch. The preprocessing branch estimates inter-frame optical flow and sharp maps to provide prior information for subsequent deblurring. In contrast, the feature reconstruction branch employs a wavelet transform-based design to recover structure and detail information from low-frequency and high-frequency sub-bands, respectively.

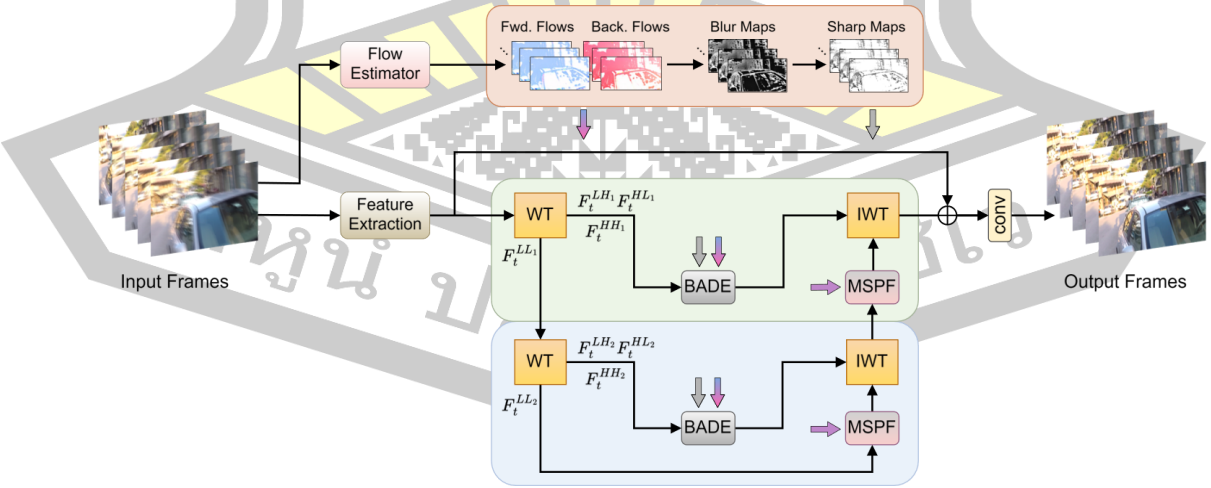


Figure 4.1 The architecture of the proposed wavelet-based blur-aware decoupled network (WBDNet).

In the preprocessing branch, we first utilize an optical flow estimation network [98] to compute motion information between adjacent frames in a video sequence. Specifically, for any frame I_i^{Blur}, we calculate the optical flow $\{O_{i\to t}\}_{t=1,t\neq i}^{N}$ between this frame and others. Considering that motion is the primary cause of blur, with higher motion corresponding to increased blur, we use optical flow between adjacent frames to estimate the blur map B_i for video frame I_i^{Blur}:

$$B_i = (O_{i\to i+1})^2 + (O_{i\to i-1})^2 \tag{4.1}$$

For the first and last frames of the video sequence, we set $B_1 = 2(O_{1\to 2})^2$ and $B_N = 2(O_{N\to N-1})^2$, respectively. Subsequently, we normalize the blur map B_i to the $[0,1]$ interval to obtain the normalized blur map B_i' and corresponding sharp map S_i:

$$B_i' = \frac{B_i - min(B_i)}{max(B_i) - min(B_i)} \tag{4.2}$$

$$S_i = 1 - B_i' \tag{4.3}$$

In the feature reconstruction branch, we first extract shallow features $\{F_t\}_{t=1}^{N}$ from video frames through several residual blocks. We then use wavelet transform to decompose these features into one low-frequency sub-band $\{F_t^{LL}\}_{t=1}^{N}$ and three high-frequency sub-bands $\{F_t^{HL}\}_{t=1}^{N}$, $\{F_t^{LH}\}_{t=1}^{N}$ and $\{F_t^{HH}\}_{t=1}^{N}$. To achieve multi-scale feature representation, we further decompose the low-frequency sub-band $\{F_t^{LL}\}_{t=1}^{N}$ into corresponding low and high-frequency sub-bands.

We design specialized processing modules based on the characteristics of different frequency band features. For low-frequency features, we develop the multi-scale progressive fusion (MSPF) module to handle structural reconstruction. For high-frequency features, we create the blur-aware detail enhancement (BADE) module to focus on detail recovery. This decoupled design enables the network to process frame features at different scales, more specifically.

4.2.2 Multi-scale Progressive Fusion Module

The low-frequency sub-bands obtained through wavelet transform carry the main structural information of the frames. Unlike other video restoration tasks, video deblurring requires not only detailed recovery but also reconstruction of structure distortions caused by blur. Based on this characteristic, we designed the multi-scale progressive fusion (MSPF) module specifically to address the structural information reconstruction in low-frequency sub-band features.

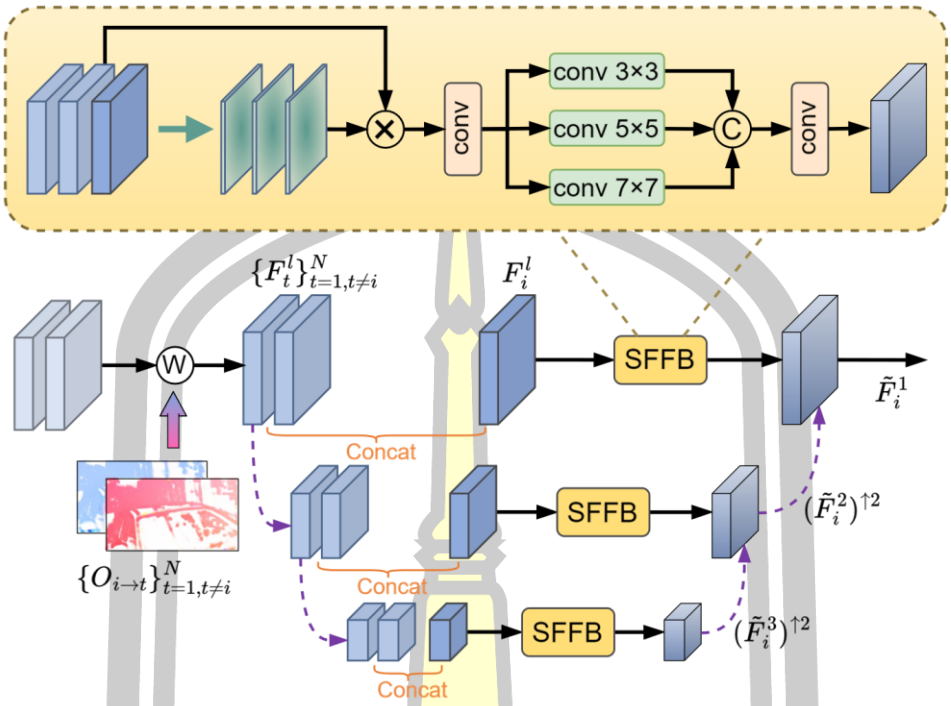


Figure 4.2 Schematic diagram of multi-scale progressive fusion (MSPF) module.

The core idea of the MSPF module is to utilize redundant structural information from multiple frames to assist in target frame restoration. As shown in Figure 4.2, we first align features from other frames to the target frame using optical flow information obtained from the preprocessing branch. Considering that low-frequency sub-bands mainly contain structural information, we adopt basic optical flow alignment instead of computationally intensive deformable convolution, maintaining effectiveness while improving computational efficiency. Denoting the target frame features as F_i^l and aligned features from other frames as $\{F_t^l\}_{t=1,t\neq i}^N$, where $l = 1$ represents the original scale, we construct a three-level feature pyramid ($l = 1,2,3$) using strided convolution filters for two 2× downsampling operations to expand the network's receptive field for capturing broader structural information.

We employ a bottom-up fusion strategy, starting from the lowest scale ($l = 3$) and proceeding level by level. At each scale level, we first concatenate the target frame features with other frame features along the channel dimension and input them into the structure feature fusion block (SFFB) for processing. The fusion process at $l = 2$ can be expressed as:

$$\tilde{F}_i^2 = \mathrm{Conv}\left(\left[H_{SFFB}([F_1^2, \ldots, F_N^2, F_i^2]), \left(\tilde{F}_i^3\right)^{\uparrow 2}\right]\right), \tag{4.4}$$

where [·, ·, ·] denotes feature concatenation operation and $(\cdot)^{\uparrow 2}$ represents 2× upsampling. This process continues progressively until output features are obtained at the original scale ($l = 1$).

In the SFFB, to emphasize common structural features and suppress interference, we first compute similarity maps between the target frame and other frame features, reflecting structural correlation at each position through similarity distances. By multiplying features with similarity maps, we obtain features with enhanced structural information. This process is represented as:

$$M(t) = \text{Sigmoid}\left(\theta\left(F_i^l\right) \odot \phi\left(F_t^l\right)\right), t \in [1, N] \text{ and } t \neq i, \tag{4.5}$$

$$F_t^{l'} = F_t^l \odot M(t), \tag{4.6}$$

where $\theta(\cdot)$ and $\phi(\cdot)$ are feature embedding functions implemented through convolution layers; $\odot$ denotes element-wise multiplication; and the sigmoid function normalizes similarity values to the [0,1] interval. The enhanced features are concatenated along the channel dimension and compressed through a convolution layer. To obtain more comprehensive structural representations, SFFB employs parallel convolution kernels of different sizes to extract multi-receptive field features, fusing structural information at different scales for complementary enhancement. This design ensures the network can comprehensively capture and integrate structural information in both temporal and scale domains.

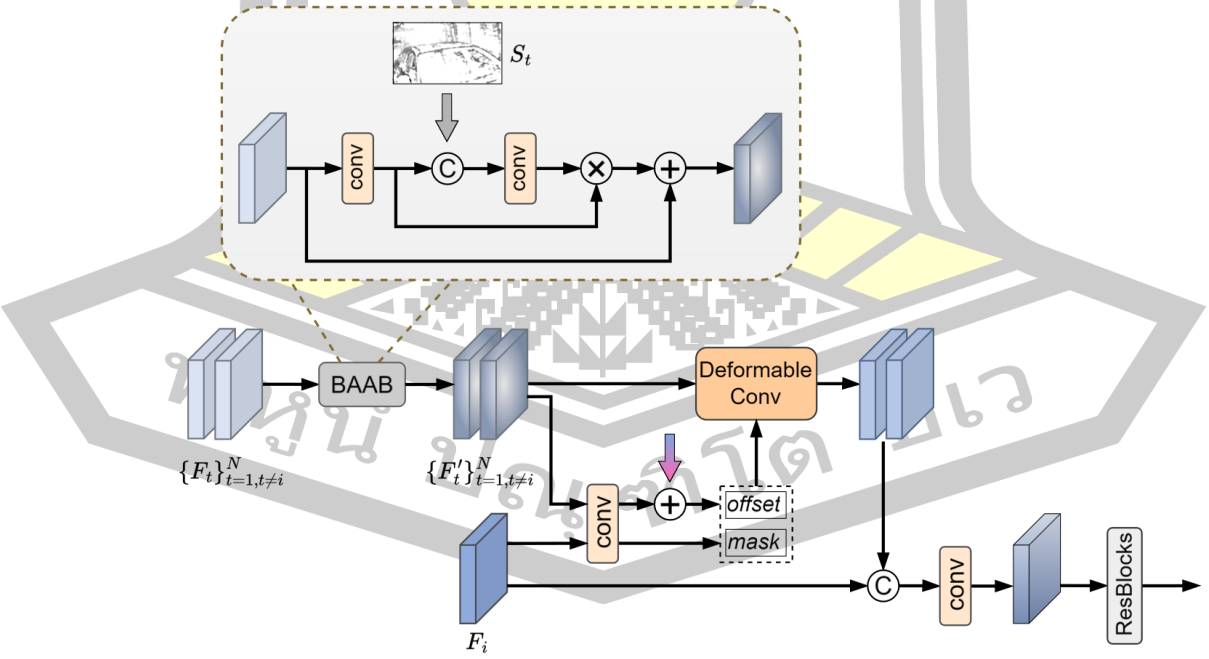


Figure 4.3 Schematic diagram of blur-aware detail enhancement (BADE) module.

4.2.3 Blur-Aware Detail Enhancement Module

High-frequency sub-bands, as crucial components of wavelet transform, contain rich, detailed information on frames. To effectively recover these details, we propose the blur-aware detail enhancement (BADE) module. The core design principle of this module is to fully utilize temporal information from video sequences while considering the impact of motion blur for precise detail recovery. We first concatenate high-frequency sub-band features $\{F_t^{HL}\}_{t=1}^{N}$, $\{F_t^{LH}\}_{t=1}^{N}$ and $\{F_t^{HH}\}_{t=1}^{N}$ along the channel dimension as input to the BADE module.

The specific details of the BADE module are shown in Figure 4.3. Let F_i denote the target frame features and $\{F_t\}_{t=1,t\neq i}^{N}$ denote features from other frames. Other frame features first pass through a blur-aware attention block (BAAB). In this block, we incorporate the sharp map S_t from the preprocessing branch. This sharp map provides crucial prior information about motion-blurred regions in the video. Using this information, we adaptively modulate inter-frame features to reduce the impact of blurred pixels. Specifically, the computation process of BAAB is:

$$F_t^{att} = \mathrm{Conv2}([\mathrm{Conv1}(F_t), S_t]), \tag{4.7}$$

$$F_t' = \mathrm{Conv1}(F_t) \odot F_t^{att} + F_t, \tag{4.8}$$

where $[\cdot,\cdot]$ represents feature concatenation operation and $\odot$ denotes element-wise multiplication. This design enables the network to dynamically adjust feature responses based on sharpness information, enhancing representations in sharp regions while suppressing the influence of blurred regions. The residual connection ensures the effective transmission of original feature information.

After obtaining enhanced feature representations, we adopt a flow residual learning approach to obtain sampling offsets for deformable convolution, which aligns $\{F_t'\}_{t=1,t\neq i}^{N}$ to F_i individually through deformable convolution:

$$\Delta P_t = O_{i\to t} + \mathcal{C}^{p}([F_t', F_i]), \tag{4.9}$$

$$\Delta M_t = \mathrm{Sigmoid}\big(\mathcal{C}^{m}([F_t', F_i])\big), \tag{4.10}$$

$$F_t^{a} = \mathrm{DConv}\big(F_t', \Delta P_t, \Delta M_t\big), \tag{4.11}$$

where $\mathcal{C}$ denotes a stack of convolutions, ΔP and ΔM represent the offset and modulation factor for deformable convolution, respectively. This design offers two key advantages. The first advantage relates to feature enhancement. Through BAAB, detail features in sharp regions become more prominent, which leads to improved feature-matching accuracy. The second advantage concerns motion handling. By

using flow residual learning to guide the offset field estimation in deformable convolution, the network can better adapt to large-scale motion changes in complex scenes. Overall, this design facilitates the effective transmission of detailed information throughout the alignment process.

Finally, we concatenate all aligned frame features along the channel dimension, compress the channel dimension through one convolution layer, and use cascaded residual blocks for deep feature extraction. It enables the network to fully integrate multi-frame information, further enhancing the expressiveness of detail features.

4.3 Dataset and Training Details

4.3.1 Dataset

1) DVD Dataset

The DVD dataset [78] consists of 71 video sequences captured using a high-speed camera at 240fps. To generate realistic motion blur, blurred frames at 30fps are synthesized by averaging consecutive sharp frames. The dataset provides paired blurred-sharp frame sequences covering diverse scenarios, including indoor and outdoor scenes, camera shake, object motion, and varying lighting conditions. The training set includes 61 sequences, with the remaining 10 sequences reserved for testing.

2) GoPro Dataset

The GoPro dataset [71] contains 33 video sequences captured using a GoPro camera at 240fps. Similar to DVD, motion blur at 30fps is simulated by averaging multiple consecutive sharp frames, resulting in 3,214 blurred-sharp frame pairs. These sequences primarily feature outdoor scenes with significant camera motion and dynamic objects, making them particularly challenging for deblurring tasks. The dataset is divided into 22 training sequences and 11 test sequences. The GoPro dataset is widely used due to its effective simulation of real motion blur characteristics encountered in consumer-grade cameras.

3) BSD Dataset

The BSD dataset [81] is a real-world video deblurring dataset, collected using a beam-splitter acquisition system with two synchronized cameras. It contains three different blur intensity configurations (1ms-8ms, 2ms-16ms, and 3ms-24ms exposure pairs), with each containing 100 video sequences. The sequences cover diverse real-world scenarios featuring various motion patterns, including camera shake and object motion in both indoor and outdoor environments. The BSD dataset has demonstrated

superior generalization capability compared to synthetic datasets, making it particularly valuable for developing and evaluating video deblurring algorithms intended for real-world applications.

4.3.2 Training Details

During the training process, we employ data augmentation to enhance the diversity of our training data. This includes techniques such as horizontal or vertical flipping, 90° rotation, and random cropping of the images. Our batch size is set to 8 to optimize the training process and the network extracts image patches of size 160×160 for training. Our model is trained by Adam optimizer [89] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-4}. All experiments were conducted on two NVIDIA RTX 2080Ti GPUs using PyTorch. We train the network end-to-end by minimizing L1 loss between the generated sharp frames and the ground truth frames.

Table 4.1 Quantitative comparisons on the DVD dataset.

Method	**DVD** [78]	**SRN** [72]	**DVDSFE** [99]	**ESTRNN** [81]	**CDVD-TSP** [8]	**MPRNet** [100]
PSNR	30.01	29.98	31.71	32.01	32.13	32.24
SSIM	0.8877	0.8842	0.9159	0.9162	0.9268	0.9253
Method	**STDAN** [10]	**RNN-MBP** [82]	**FGST** [13]	**LightViD** [101]	**STCT** [102]	**Ours**
PSNR	33.05	32.49	33.36	32.51	<u>33.45</u>	**34.28**
SSIM	0.9374	<u>0.9568</u>	0.9500	0.9460	0.9421	**0.9655**

4.4 Experimental Result and Discussion

4.4.1 Comparison with the State-of-the-art Methods

1) Evaluations on the DVD dataset

To validate the effectiveness of our proposed method, we conducted comprehensive performance evaluations on the DVD dataset. As shown in Table 4.1, our approach outperforms state-of-the-art methods in both PSNR and SSIM metrics, demonstrating its superior performance in video deblurring tasks. To further illustrate the visual quality advantages of our method, we present comparative deblurring results from different approaches in Figure 4.4. Through magnified comparisons of specific regions, the differences in detail restoration between methods become more

apparent. In the sequence "IMG_0030", the intersection sign region suffered severe structural degradation due to motion blur, and existing methods struggled to accurately restore its complete contour. In contrast, our method successfully reconstructed clear outlines of the sign through a strategy that decouples structural and detailed information processing, validating the effectiveness of this design approach. The advantages of our method are even more pronounced in the sequence "IMG_0021" comparison. Not only did it accurately restore the overall shape of the wheel, but it also precisely reconstructed fine structures such as spokes, achieving results closest to the ground truth. These outcomes further confirm that our proposed method can maintain both structural integrity and detail authenticity when processing complex scenes, demonstrating excellent deblurring capabilities.

Figure 4.4 Comparison of visual results on the DVD dataset.

2) Evaluations on the GoPro dataset

We conducted comparative evaluations on the GoPro dataset. As shown in Table 4.2, our proposed method achieved excellent objective metrics on this dataset, reaching leading performance in both PSNR and SSIM. Figure 4.5 presents visual comparison results of different methods on the GoPro dataset. Unlike the DVD dataset which primarily contains static objects, the GoPro dataset includes more moving objects and complex textures, presenting new challenges for deblurring tasks. In the sequence "GOPR0384_11_00", despite severe motion blur in the pedestrian's foot area due to rapid movement, our method successfully reconstructed clear

boundary contours. For the sequence "GOPR0385_11_01", in complex regions where pedestrian hands overlap with building edges, our method not only recovered hand details but also avoided edge confusion, demonstrating strong robustness in handling complex scenes. In the sequence "GOPR0410_11_00", while other methods could partially restore the basic outlines of numbers, our method reconstructed sharper and clearer digits that more closely match the visual quality of the original image.

Table 4.2 Quantitative comparisons on the GoPro dataset.

Method	**DVD** [78]	**SRN** [72]	**DVDSFE** [99]	**ESTRNN** [81]	**CDVD-TSP** [8]	**MPRNet** [100]
PSNR	27.31	30.29	31.01	31.07	31.67	32.73
SSIM	0.8255	0.9014	0.9130	0.9023	0.9279	0.9366
Method	**STDAN** [10]	**RNN-MBP** [82]	**FGST** [13]	**LightViD** [101]	**STCT** [102]	**Ours**
PSNR	32.62	<u>33.32</u>	32.90	32.73	32.97	**34.87**
SSIM	0.9375	<u>0.9627</u>	0.9610	0.9410	0.9406	**0.9729**

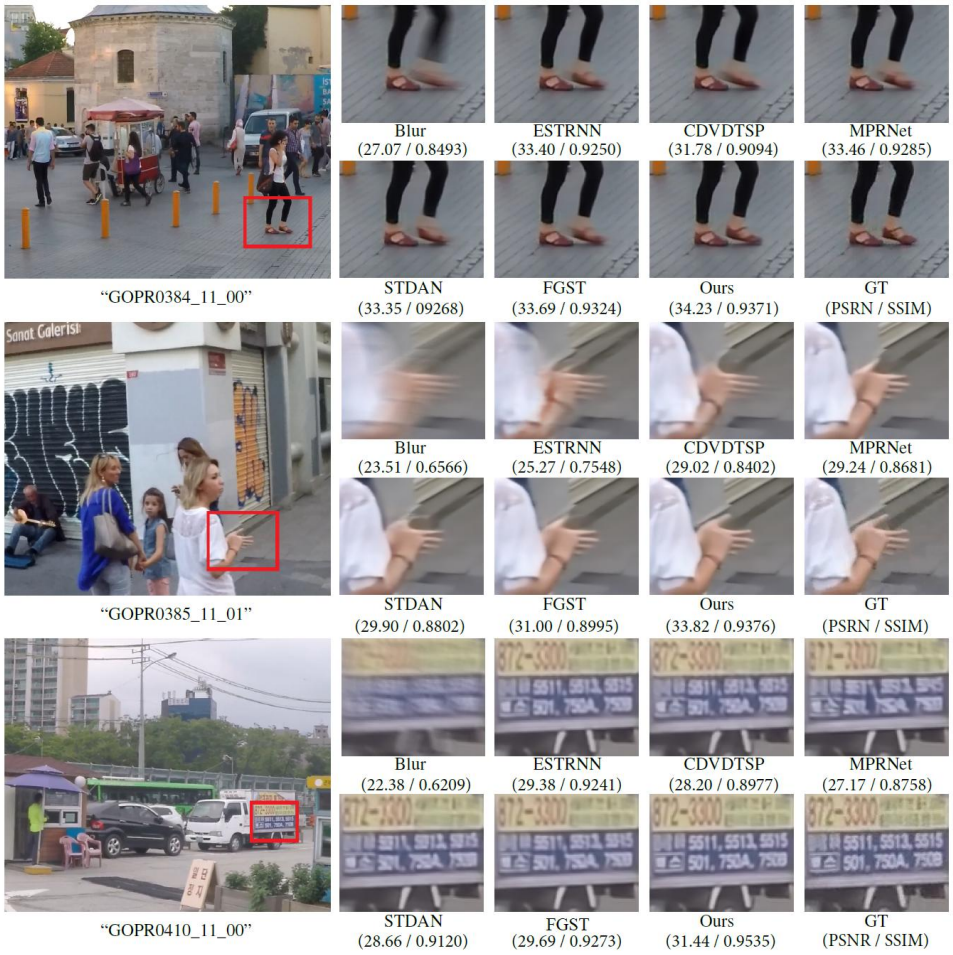


Figure 4.5 Comparison of visual results on the GoPro dataset.

3) Evaluations on the BSD dataset

To further validate our method's generalization capability, we evaluate its performance on the BSD dataset. This dataset features diverse blur characteristics through different exposure time configurations. Longer exposure times (3ms-24ms) lead to more severe blur, posing greater challenges for deblurring algorithms. As shown in Table 4.3, our method achieves superior performance across all exposure configurations.

Figure 4.6 demonstrates our deblurring results under three typical motion patterns. In static camera scenes, we successfully restore sharp contours of blurred motorcyclists. The original clear background details are also well preserved. For scenes with aligned camera and object motion, our method shows excellent performance. It accurately reconstructs the car front's edge structure and precisely recovers the front grille's texture. In scenes with opposing camera and object motion, our method effectively recovers vehicle body text. The restored text maintains high clarity without introducing artifacts. These results demonstrate our method's robustness and superiority in handling real-world motion blur patterns.

Table 4.3 Quantitative comparisons on the BSD dataset.

Method	**1ms-8ms**		**2ms-16ms**		**3ms-24ms**	
	PSNR	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
SRN [72]	31.84	0.917	29.95	0.891	28.92	0.882
STFAN [103]	32.78	0.922	32.19	0.919	29.47	0.872
CDVD-TSP [8]	33.54	0.942	32.16	0.926	31.58	0.926
ESTRNN [81]	33.36	0.937	31.95	0.925	31.39	0.926
STDAN [10]	34.32	0.946	33.27	0.942	32.83	0.944
FGST [13]	<u>34.60</u>	<u>0.969</u>	<u>33.62</u>	<u>0.962</u>	<u>33.16</u>	<u>0.959</u>
Ours	**34.85**	**0.971**	**33.91**	**0.969**	**33.34**	**0.966**

Figure 4.6 Comparison of visual results on the BSD dataset (3ms-24ms).

4) Network Efficiency Analysis

Table 4.4 shows the comparison results in terms of parameters, inference time, computational complexity and restoration quality for processing one 1280×720 resolution video frame in GoPro dataset. The experimental results demonstrate that our method achieves the best restoration quality while maintaining moderate computational costs. In terms of model efficiency, our method employs 8.9M parameters, which is higher than ESTRNN (2.47M) and STFAN (5.37M) but significantly lower than CDVD-TSP (16.2M) and STDAN (13.8M). Similarly, while our computational complexity is higher than ESTRNN and STFAN, it remains more efficient than CDVD-TSP and STDAN. For processing speed, our method can process one frame in 0.41 seconds, ranking in the middle among all compared methods, which demonstrates a good balance between restoration quality and computational efficiency.

Table 4.4 Model parameters, inference time and computational cost comparisons on the GoPro dataset.

Method	**STFAN** [103]	**CDVD-TSP** [8]	**ESTRNN** [81]	**STDAN** [10]	**FGST** [13]	**Ours**
Para. (M)	5.37	16.2	2.47	13.8	9.7	8.9
Time (s)	0.16	2.03	0.21	3.24	0.78	0.41
GMACs	144.26	1275.71	203.74	1524.22	579.87	746.53
PSNR	28.59	31.67	31.07	32.62	32.90	34.87

4.4.2 Ablation Study

To gain deeper insights into the working mechanism of our proposed method and validate the effectiveness of core components, we conducted systematic ablation

studies. Our network design is based on a key concept: decomposing the deblurring task into two subtasks through wavelet transform. The first subtask focuses on structure restoration, handled by the multi-scale progressive fusion (MSPF) module. The second subtask addresses detail enhancement, managed by the blur-aware detail enhancement (BADE) module.

Table 4.5 Ablation study of the proposed module on the GoPro dataset.

Exp.	**(a)**	**(b)**	**(c)**
MSPF		✓	✓
BADE	✓		✓
PSNR	31.47	29.88	**34.87**
SSIM	0.9215	0.8965	**0.9729**

Figure 4.7 Effectiveness of the proposed module for video deblurring.

1) Effectiveness of the MSPF and BADE module

We first validate the necessity of our wavelet-based task decomposition design through ablation studies. As shown in Table 4.5, removing the MSPF module leads to a 3.4dB drop in PSNR. Similarly, removing the BADE module results in a 4.99dB decrease. These results demonstrate that both modules are crucial for effective deblurring.

To better understand the role of each module, we present visual comparison results in Figure 4.7. Without the MSPF module, the billboard's edge structure shows obvious distortion and blur. This is because the network lacks a dedicated module for processing low-frequency structural information. Without the BADE module, the text details on the billboard become completely indiscernible. This verifies the importance of high-frequency detail enhancement. Only with both modules working together can the network achieve optimal restoration results. This synergy enables accurate reconstruction of both structural integrity and fine details.

Table 4.6 Ablation study for MSPF module design on the GoPro dataset.

MSPF module			
Progressive Fusion		✓	✓
SFFB	✓		✓
PSNR	33.15	32.26	**34.87**
SSIM	0.9621	0.9345	**0.9729**

w/o Progressive Fusion w/o SFFB w/ PF, w/ SFFB GT

Figure 4.8 Visualization of ablation on MSPF module.

2) Effectiveness of components in the MSPF module

We evaluated the effectiveness of the MSPF module design. This module utilizes redundant structural information across multiple frames through a progressive fusion strategy and structure feature fusion block (SFFB). Experimental results in Table 4.6 show that removing the progressive fusion strategy and SFFB leads to PSNR decreases of 1.72dB and 2.61dB, respectively.

The visual comparison in Figure 4.8 demonstrates the effectiveness of two key components in the MSPF module. Without the progressive fusion strategy, the vehicle's overall outline appears noticeably blurred. This highlights the importance of this strategy in structure reconstruction. When removing the SFFB module, the vehicle's contours show motion trailing effects. It indicates that efficient multi-frame temporal information fusion is crucial for achieving ideal deblurring results. With both components working together, the reconstruction results achieve the best visual quality. These results validate the effectiveness of our MSPF module design.

Table 4.7 Ablation study for BADE module design on the GoPro dataset.

BADE module			
BAAB		✓	✓
DConv Align	✓		✓
PSNR	33.06	30.73	**34.87**
SSIM	0.9616	0.9085	**0.9729**

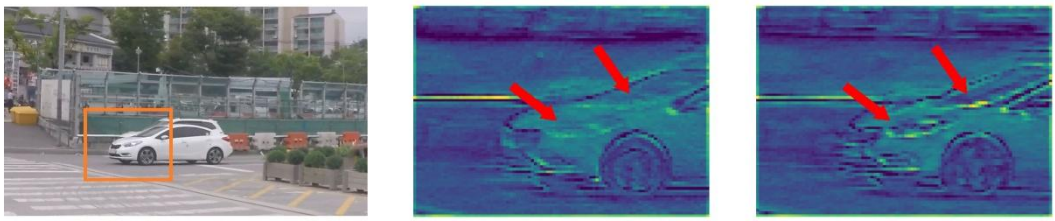


Figure 4.9 Feature maps visualization of ablation on blur-aware attention block (BAAB).

3) Effectiveness of components in the BADE module

We conduct in-depth ablation analysis on two components of the BADE module: the blur-aware attention block (BAAB) using sharpness priors and deformable convolution alignment based on optical flow residuals. As shown in Table 4.5 and Table 4.7, adding BAAB alone yields limited improvement (0.85dB). However, incorporating BAAB on top of deformable convolution alignment shows more significant improvement (1.81dB). This reveals important synergistic effects between BAAB and feature alignment operations.

To better understand BAAB's working mechanism, we visualize intermediate feature maps in Figure 4.9. Without BAAB, the vehicle headlight area suffers more interference from blurred pixels. This leads to decreased detail restoration quality. BAAB adaptively modulates feature responses using sharpness prior information. It enhances feature responses in sharp regions while suppressing them in blurred areas. The combination of BAAB and deformable convolution alignment maximizes their respective strengths. BAAB provides sharpness awareness to help identify and preserve features in clear regions. Meanwhile, deformable convolution ensures effective temporal propagation and fusion of these valuable features. This synergy enables better detail recovery in complex dynamic scenes.

4.4.3 Limitations

In testing across two challenging scenarios, our method demonstrates strengths while revealing areas for improvement, which can be seen in Figure 4.10. In the left case, for high-speed moving vehicles, the method preserves basic structure but shows limitations in reconstructing precise edge details. In the right case, when processing complex crowd scenes with multiple subjects, the method successfully captures the overall composition but falls short in contour clarity and detail recovery for the raised hands of central figures. These cases highlight the inherent technical difficulties in video deblurring, particularly in extreme motion conditions and complex scenes, where balancing structure preservation and detail reconstruction remains an

unresolved challenge. Nevertheless, the proposed method establishes a reliable foundation in overall structure recovery and visual consistency, providing clear directions for future improvements in high-frequency detail preservation and complex motion scene processing.

Figure 4.10 The failure cases of the WBDNet.

4.5 Conclusion

In this chapter, we propose a novel wavelet-based blur-aware decoupled network (WBDNet). Through wavelet transform, we decompose the deblurring task into two subtasks: structure restoration and detail enhancement. For the low-frequency domain, we design a multi-scale progressive fusion (MSPF) module that effectively integrates multi-frame structural information through a multi-scale feature fusion strategy and structure feature fusion block. For the high-frequency domain, we introduce a blur-aware detail enhancement (BADE) module that combines blur-aware attention block utilizing sharpness priors and deformable convolution alignment based on optical flow residuals to enhance detail reconstruction. Experimental results on benchmark datasets demonstrate that our proposed method achieves excellent performance in both objective metrics and subjective visual quality.

Chapter 5

Conclusions and Future Work

5.1 Summary and Discussion

This research addresses two critical challenges in video restoration: space-time video super-resolution and video deblurring. Through systematic research and extensive experiments, we have made the following key contributions:

1) For space-time video super-resolution:

- We proposed a novel deformable attention network (DANet) that effectively handles both spatial and temporal super-resolution in a unified framework.
- The designed deformable interpolation block (DIB) significantly improves intermediate frame generation by enhancing inter-frame motion capture capabilities.
- The temporal fusion module, consisting of the temporal feature shuffle block (TFSB) and motion feature enhancement block (MFEB), enables effective exploitation of complementary temporal information across multiple frames.

2) For video deblurring:

- We developed a wavelet-based blur-aware decoupled network (WBDNet) that effectively decouples structure recovery from detail enhancement.
- The multi-scale progressive fusion (MSPF) module successfully reconstructs structural information through hierarchical feature fusion and multi-frame information integration.
- The blur-aware detail enhancement (BADE) module combines sharpness priors with advanced alignment techniques to achieve superior detail restoration.

Extensive experiments on multiple benchmark datasets demonstrate that the proposed novel architectures and methodologies effectively address the challenges of joint spatial-temporal enhancement and complex blur removal, outperforming existing methods in terms of objective metrics and visual quality, while maintaining computational efficiency. Overall, this research significantly advances the field of video restoration and offers strong practical value for diverse applications ranging from multimedia entertainment to surveillance systems. Furthermore, the modular design facilitates adaptation to specific domain requirements, positioning this work as

a valuable foundation for future research and development in video enhancement technologies.

5.2 Future Work

Despite the advances made in this thesis, several important research directions remain to be explored. One promising direction is the development of unified frameworks capable of jointly handling multiple degradation factors. This would enable more efficient processing of real-world videos that often suffer from multiple quality issues simultaneously, such as low resolution, motion blur, and compression artifacts.

Further research could focus on enhancing temporal information utilization through more advanced feature alignment and fusion strategies. This includes developing methods to exploit long-range temporal dependencies while maintaining computational efficiency, and investigating adaptive temporal window selection mechanisms for optimal information utilization. Such advances would be particularly beneficial for processing videos with complex motion patterns.

From a practical perspective, future work should address the challenges of real-world deployment. This includes improving model efficiency for real-time processing on resource-constrained devices, developing more robust methods for handling diverse real-world degradation patterns, and investigating unsupervised or self-supervised learning approaches to reduce dependence on paired training data.

REFERENCES

[1] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, 2019.

[2] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "Mucan: Multi-correspondence aggregation network for video super-resolution," presented at the Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part X 16, 2020, 2020.

[3] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 2021.

[4] M. Haris, G. Shakhnarovich, and N. Ukita, "Space-time-aware multi-resolution video enhancement," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2020.

[5] S. Y. Kim, 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, vol. 34, no. 07, pp. 11278-11286.

[6] J. Wulff and M. J. Black, "Modeling blurred video with layers," presented at the Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, 2014, 2014.

[7] T. Hyun Kim and K. Mu Lee, "Generalized video deblurring for dynamic scenes," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 2015.

[8] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2020.

[9] J. Liang *et al.*, "Recurrent video restoration transformer with guided deformable attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 378-393, 2022 2022.

[10] H. Zhang, H. Xie, and H. Yao, "Spatio-temporal deformable attention network for video deblurring," presented at the European Conference on Computer Vision, 2022, 2022.

[11] B. Jiang, Z. Xie, Z. Xia, S. Li, and S. Liu, "Erdn: Equivalent receptive field deformable network for video deblurring," presented at the European Conference on Computer Vision, 2022, 2022.

[12] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "Vdtr: Video deblurring with transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 160-171, 2022 2022.

[13] J. Lin *et al.*, "Flow-Guided Sparse Transformer for Video Deblurring," presented at the International Conference on Machine Learning, 2022, 2022.

[14] J. Liang *et al.*, "Vrt: A video restoration transformer," *IEEE Transactions on Image Processing*, 2024 2024.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and*

pattern recognition, 2016, pp. 770-778.

[16] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136-144.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.

[19] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng, "Efficient spatio-temporal recurrent neural network for video deblurring," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 2020: Springer, pp. 191-207.

[20] S. Mehta *et al.*, "Evrnet: Efficient video restoration on edge devices," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 983-992.

[21] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," presented at the Proceedings of the European conference on computer vision (ECCV), 2018, 2018.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3-19.

[23] M. Zhao, Y. Xu, and S. Zhou, "Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5646-5654.

[24] T. Isobe *et al.*, "Video super-resolution with temporal group attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8008-8017.

[25] J. Cao, Y. Li, K. Zhang, and L. Van Gool, "Video super-resolution transformer," *arXiv preprint arXiv:2106. 06847*, 2021 2021.

[26] J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764-773.

[27] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, 2019.

[28] Y. Tian, Y. Zhang, Y. Fu, and C. X. Tdan, "temporally-deformable alignment network for video super-resolution. In 2020 IEEE," presented at the CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 2020.

[29] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," presented at the Proceedings of the IEEE international conference on computer vision, 2017, 2017.

[30] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," presented at the Proceedings of the IEEE conference on computer

vision and pattern recognition, 2018, 2018.
[31] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2018.
[32] L. Yuan, Y. Chen, H. Liu, T. Kong, and J. Shi, "Zoom-in-to-check: Boosting video interpolation via instance-level discrimination," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 2019.
[33] J. Park, K. Ko, C. Lee, and C.-S. Kim, "Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation," presented at the Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XIV 16, 2020, 2020.
[34] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 2020.
[35] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2017.
[36] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," presented at the Proceedings of the IEEE international conference on computer vision, 2017, 2017.
[37] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 2019.
[38] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 933-948, 2019 2019.
[39] H. Lee, T. Kim, T.-Y. Chung, D. Pak, Y. Ban, and S. Lee, "Adacof: Adaptive collaboration of flows for video frame interpolation," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 2020.
[40] X. Cheng and Z. Chen, "Video frame interpolation via deformable separable convolution," presented at the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 2020.
[41] Z. Shi, X. Liu, K. Shi, L. Dai, and J. Chen, "Video frame interpolation via generalized deformable convolution," *IEEE transactions on multimedia*, vol. 24, pp. 426-439, 2021 2021.
[42] X. Cheng and Z. Chen, "Multiple video frame interpolation via enhanced deformable separable convolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7029-7045, 2021 2021.
[43] T. Ding, L. Liang, Z. Zhu, and I. Zharkov, "Cdfi: Compression-driven network design for frame interpolation," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 2021.
[44] J. Caballero *et al.*, "Real-time video super-resolution with spatio-temporal networks and motion compensation," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2017.

[45] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, 2017.

[46] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, pp. 1-20, 2017 2017.

[47] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 2018.

[48] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 2022.

[49] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 2018.

[50] S. Li, F. He, B. Du, L. Zhang, Y. Xu, and D. Tao, "Fast spatio-temporal residual network for video super-resolution," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 2019.

[51] S. Y. Kim, J. Lim, T. Na, and M. Kim, "Video super-resolution based on 3D-CNNS with consideration of scene change," presented at the 2019 IEEE International Conference on Image Processing (ICIP), 2019, 2019.

[52] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 1015-1028, 2017 2017.

[53] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, "Residual invertible spatio-temporal network for video super-resolution," presented at the Proceedings of the AAAI conference on artificial intelligence, 2019, 2019.

[54] D. Fuoli, S. Gu, and R. Timofte, "Efficient video super-resolution through recurrent latent space propagation," presented at the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, 2019.

[55] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2020.

[56] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015 2015.

[57] J. Cao *et al.*, "Towards interpretable video super-resolution via alternating optimization," presented at the European Conference on Computer Vision, 2022, 2022.

[58] Z. Chen *et al.*, "Videoinr: Learning video implicit neural representation for continuous space-time super-resolution," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022,

2022.
[59] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "Rstt: Real-time spatial temporal transformer for space-time video super-resolution," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 2022.
[60] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Efficient marginal likelihood optimization in blind deconvolution," presented at the CVPR 2011, 2011, 2011.
[61] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," presented at the IEEE international conference on computational photography (ICCP), 2013, 2013.
[62] W. Ren, X. Cao, J. Pan, X. Guo, W. Zuo, and M.-H. Yang, "Image deblurring via enhanced low-rank prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3426-3437, 2016 2016.
[63] T. Hyun Kim, B. Ahn, and K. Mu Lee, "Dynamic scene deblurring," presented at the Proceedings of the IEEE international conference on computer vision, 2013, 2013.
[64] J. Pan, Z. Hu, Z. Su, H.-Y. Lee, and M.-H. Yang, "Soft-segmentation guided object motion deblurring," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 2016.
[65] S. Zheng, L. Xu, and J. Jia, "Forward motion deblurring," presented at the Proceedings of the IEEE international conference on computer vision, 2013, 2013.
[66] D. Gong *et al.*, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017.
[67] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, 2015.
[68] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017.
[69] J. Zhang *et al.*, "Dynamic scene deblurring using spatially variant recurrent neural networks," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2018.
[70] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2018.
[71] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017.
[72] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2018.
[73] H. Chen *et al.*, "Pre-trained image processing transformer," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern

recognition, 2021, 2021.
[74] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 2022.
[75] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, "Full-frame video stabilization with motion inpainting," *IEEE Transactions on pattern analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1150-1163, 2006 2006.
[76] S. Cho, J. Wang, and S. Lee, "Video deblurring for hand-held cameras using patch-based synthesis," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1-9, 2012 2012.
[77] T. Hyun Kim and K. Mu Lee, "Segmentation-free dynamic scene deblurring," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, 2014.
[78] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017.
[79] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 291-301, 2018 2018.
[80] T. Hyun Kim, K. Mu Lee, B. Scholkopf, and M. Hirsch, "Online video deblurring via dynamic temporal blending network," presented at the Proceedings of the IEEE international conference on computer vision, 2017, 2017.
[81] Z. Zhong, Y. Gao, Y. Zheng, B. Zheng, and I. Sato, "Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 284-301, 2023 2023.
[82] C. Zhu *et al.*, "Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring," presented at the Proceedings of the AAAI conference on artificial intelligence, 2022, 2022.
[83] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2020.
[84] D. Li *et al.*, "Arvo: Learning all-range volumetric correspondence for video deblurring," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 2021.
[85] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017 2017.
[86] G. Xu, J. Xu, Z. Li, L. Wang, X. Sun, and M.-M. Cheng, "Temporal modulation network for controllable space-time video super-resolution," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 2021.
[87] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 346-360, 2013 2013.

[88] S. Nah *et al.*, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, 2019.
[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412. 6980*, 2014 2014.
[90] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2021, 2021.
[91] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent Back-Projection Network for Video Super-Resolution," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, 2019.
[92] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335-350, 2009 2009.
[93] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 2018.
[94] J. Gast and S. Roth, "Deep video deblurring: The devil is in the details," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, 2019.
[95] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, "Sdwnet: A straight dilated network with wavelet transformation for image deblurring," presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2021, 2021.
[96] J. Dong, J. Pan, Z. Yang, and J. Tang, "Multi-scale residual low-pass filter network for image deblurring," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, 2023.
[97] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang, "Deep discriminative spatial and temporal network for efficient video deblurring," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 2023.
[98] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017.
[99] X. Xiang, H. Wei, and J. Pan, "Deep video deblurring using sharpness features from exemplars," *IEEE Transactions on Image Processing*, vol. 29, pp. 8976-8987, 2020 2020.
[100] S. W. Zamir *et al.*, "Multi-stage progressive image restoration," presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 2021.
[101] L. Lin, G. Wei, K. Liu, W. Feng, and T. Zhao, "LightViD: Efficient Video Deblurring with Spatial-Temporal Feature Fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024 2024.
[102] L. Zhang, B. Xu, Z. Yang, and J. Pan, "Deblurring Videos Using Spatial-Temporal Contextual Transformer With Feature Propagation," *IEEE Transactions on Image Processing*, 2024 2024.
[103] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter

adaptive network for video deblurring," presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2019, 2019.

BIOGRAPHY

NAME	Hua Wang	
DATE OF BIRTH	12/10/1993	
PLACE OF BIRTH	China	
ADDRESS	Putian City, Fujian Province, China	
POSITION	Full-Time Lecture	
PLACE OF WORK	Putian University of China	
EDUCATION	2016	Bachelor in Chemical Engineering and Technology, East China University of Science and Technology
	2021	Master in Software Engineering, South China University of Technology
	2025	Doctor of Philosophy in Computer Science, Mahasarakham University
Research grants & awards	-	
Research output	-	