Selective Fusion for Video Super-Resolution with Recurrent Architecture

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

July 18th, 2020

Zichen GONG
(5118FG17-2)
Advisor: Prof. Hiroshi Watanabe
Research guidance: Research on Audiovisual Information Processing

Acknowledgements

To begin with, I would like to express my deepest gratitude to my research supervisor Prof.Hiroshi Watanabe for his sparing no effort to continuously support me in different aspects and the excellent laboratory circumstances he built for us. His academic guidance and amiable treatment are invaluable to me in my two-year master's study.

I am grateful to the joint research group members of Sharp Corporation, Mr.Takeshi Chujoh, Mr.Norio Ito, Mr.Tomohiro Ikai, and Mr.Eiichi Sasaki for the professional research advices and cutting-edge academic inspirations they provided.

I also appreciate innumerable warm helps from all of lab members, their kindness and vigor keep motivating me to become better.

Lastly, I am extremely grateful for my parents' energetic encouragements and full support to my overseas study.

Abstract

As an important subtask of video restoration, video super-resolution has attracted a lot of attention in the community as it can eventually promote a wide range of technologies, e.g., video quality enhancement, video compression, highly efficient video transmission system etc. Recent video super-resolution model with recurrent architecture achieves cutting-edge performance. It efficiently utilizes recurrent architecture with neural networks to gradually aggregate details from previously generated frames.

Nevertheless, this method faces a serious drawback that it is sensitive to occlusion, blur, and large motion changes since it only takes the previously generated output as recurrent input for the super resolution model. This will lead to undesirable rapid information loss during the recurrently generating process, and performance will therefore be dramatically decreased. Our works focus on addressing the issue of rapid information loss in video super-resolution model with recurrent architecture. By producing attention maps through selective fusion module, the recurrent model can adaptively aggregate necessary details across all previously generated high-resolution (HR) frames according to their informativeness. The proposed method is demonstrated to be useful for preserving high frequency details collected progressively from each frame, while enable the model to discarding undesired noisy artifacts that wrongly and sequentially enhanced during the recurrent super-resolution process. This significantly improves the quality of the super resolution video.

Keywords: Video super-resolution, selective fusion, video transmission system, recurrent networks

List of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 4

Chapter 1 Introduction

1.1 Video super-resolution with deep learning

Super-resolution (SR) aims at transferring low-resolution (LR) inputs to corresponding high-resolution (HR) outputs. It is an inherently ill-posed problem since one single given input in the LR space can be mapped to multiple possible outputs in the HR space. Since HR images and videos contain more high-frequency spatial details, SR is widely applied in different fields, e.g. video quality enhancement, video compression, and video transmission system etc.

According to the processing number of inputs, SR can be mainly classified into single image super-resolution (SISR) and multi-image super-resolution (MISR). Video super-resolution can be achieved by repeating the process of SISR (or MISR) and inferring every frame of a given video. Although SISR often shows an efficiently better performance than MISR, MISR is more frequently adopted in video super-resolution (VSR) task since it can utilize the temporal relevance between consecutive input LR frames, and it would be naturally prone to generate temporally consistent frames compared to conducting VSR in the SISR manner.

Thanks to the rapid development of deep learning techniques in recent years, deep learning based SR outperformed traditional methods such as interpolation-based methods [1], [2], reconstruction-based methods [3], [4] and example-based methods [5], [6], etc. Deep learning based video super-resolution model learns the mapping between LR space and HR space in an end-to-end manner. Deep neural networks would automatically extract features and abstractions during the end-to-end training procedure in the SR task. It better solved the problem of indistinct definition of the mapping among highly sophisticated dataset than in traditional methods. Consequently, most of recent state-of-the-art SR models are based on deep learning techniques.

Recent VSR work highlights using the temporal relevance by either taking multiple LR frames as inputs to generate successive HR frames (MISR) such as Kappeler et al. [7] and Tao et al. [8] (sometimes the number of input frames and scale are adaptive [9]) or adopting recurrent architecture to gradually aggregate necessary high frequency spatial details from previous generated frames such like FRVSR [10] and Haris et al. [11].

In this thesis, we mainly explored the latter scheme for the reason that: with the advantage of re-using high-frequency details, it shows superiority over the methods which taking multiple frames to do inferring in both efficiency and accuracy, and showing more competitive performance in VSR quality.

1.2 Problem statement

FRVSR [10] is a typical model adopting recurrent architecture for VSR. It is an end-to-end trainable VSR framework that takes the previously estimated output as the input for the next inference. The advantage of such recurrent architecture based VSR model is shown in two aspects, one is avoiding computational redundancy since it is able to re-use the high-frequency details during the inferring process, and the other is that the output frames are naturally of temporal consistency between neighboring frames with satisfying quality.

However, it faces a serious drawback of rapid losing prior high-frequency details that collected from previously generated HR frames, because it merely and inflexibly takes one previously estimated HR frame as the recurrent input. As a result, it is sensitive to occlusion, blur, and motion changes, and the performance will therefore be largely limited.

Moreover, the recurrent architecture based VSR with inflexible input scheme is unable to discard some unnecessary high-frequency details that previously produced, especially when the scene lasts for a long sequence of frames. These unsatisfactory details will be gradually reinforced during the recurrent inference process and further degrade the overall performance as demonstrated in Figure 1.1.

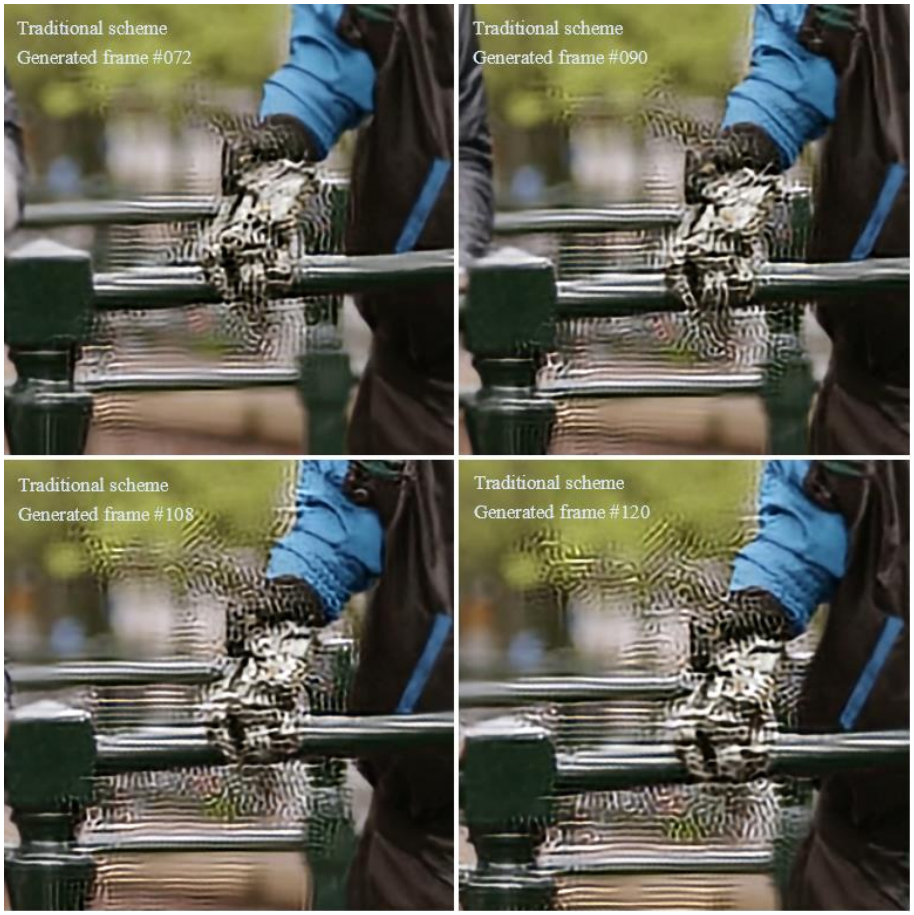


Figure 1. 1. Gradually reinforced undesirable details become noisy artifacts that degrade the overall performance (picture source: [31])

In this work, we focus on addressing the aforementioned issue by introducing a selective fusion module for appropriate locating and fusing necessary high-frequency spatial information to better reconstruct HR outputs.

1.3 Thesis outline

The outline of this thesis is organized as follows:

Chapter 1: We describe the background of VSR with deep learning and the problem that needed to be solved in this work. Besides, the advantages and disadvantages of recurrent architecture based VSR are also presented in this chapter.

Chapter 2: We introduce the technologies related to this work, ranging from the classification of SR, the principle knowledge of CNN and the key issues concerning with the recurrent architecture based VSR models to the quality assessment metrics of SR results. Through analyzing the contributions, focuses and the limitations of previous recurrent architecture based VSR work, the potential benefits of our work have been shown.

Chapter 3: We demonstrate the frameworks of the proposed selective fusion based VSR model with recurrent architecture. And we respectively introduce the details of the three stages: motion alignment, selective fusion and reconstruction stage for generating the current HR estimate, in which we explain the design principle and assumption of the selective fusion module. Besides, the reason leading to a very limited additional computation cost of the proposed method has been discussed in this chapter.

Chapter 4: The experimental environment is introduced in this chapter. By training and inferring the recurrent architecture based VSR models with different settings on test dataset, we compare and analyze the evaluation results quantitatively, along with the illustrated qualitative results, we demonstrate the superiority of our proposed method and the effectiveness of the fusion principle.

Chapter 5: Chapter 5 concludes this thesis.

Chapter 2 Related Technologies

2.1 Super-resolution categories

Super-resolution could mainly be classified into three categories, including interpolation-based methods, reconstruction-based methods, and example-based methods.

2.1.1 Interpolation-based methods

Image interpolation is a widely used image processing technology that aims to resize digital images, and it is also the simplest and the most straightforward form of conducting super-resolution (image upsampling).

Although it is computationally efficient but of low accuracy compared to other super-resolution methods, some of them are still adopted and performed an important role in deep learning based SR. Well-known interpolation-based methods include bicubic interpolation [1] and Lanczos resampling [2]. Bicubic interpolation is widely applied in building SR datasets by degrading HR images to their LR counterparts.

There is a noticeable characteristic of interpolation-based methods that they could only resize the image resolution by interpolating with known image signals other than bringing extra information, which means they could not bring high frequency details that lead a LR image to a truly HR image.

2.1.2 Reconstruction-based methods

As the principle of super-resolution is to explore the mapping solutions between LR space and HR space, reconstruction-based SR methods such as [3], [4], [12] tried to utilize sophisticated prior knowledge to restrict the possible mapping solutions so that increasing the mapping accuracy. But reconstruction-based methods often suffer from the difficulty of applying to large datasets since large amount of similar image patches are needed and they are computationally expensive.

2.1.3 Traditional learning-based methods

Learning-based SR methods are also known as example-based methods. These methods take advantage of machine learning algorithms to learn the mapping solutions between the LR space

and HR space by analyzing the statistical relationships of training pairs such as the Markov random field based approach [5] and the random forest based approach [6].

It is notable that most previously researched learning-based SR methods often suffer from the shortcoming of machine learning techniques that it is necessary to formulate a series of robust handcrafted features which are suitable for a massive dataset. It would lead to a poor definition of the mapping solutions between the LR space and the HR space.

2.1.4 Recent deep learning based method

Deep learning models is able to extract the hierarchical features of images automatically in an end-to-end manner and then leverage them to achieve the purpose defined by objective functions (loss functions). Because of the superior robustness of these learning-based extracted abstractions, the highly efficient end-to-end learning process, and the development of hardware computing power, deep learning based SR models achieved state-of-the-art performance.

2.2 Convolutional Neural Network

Convolutional Neural Network (CNN) was the first introduced deep learning based model for SR by Dong et al. [13]. And it achieved state-of-the-art performance. Many subsequent researchers followed up with deeper neural networks of different types of architectures, loss functions and learning strategies, carrying the field into a new era.

CNN is originally developed for image classification task. It is made up of neurons that have learnable weights and biases, which is the analogy of biological brain neurons. Every neuron receives some inputs and performs a dot product with an optionally followed non-linearity transformation. A CNN consists of a sequence of layers including convolutional layer, pooling layer and fully-connected layer. They transform one volume of activations (feature map) to another through a differentiable function. And the whole network will express a single differentiable score function for classification problem.

The neurons in a layer of the CNN are arranged in 3 dimensions: width, height, and depth. They are of an important characteristic: local connectivity, it means that the different layers of neurons in CNN connect to only a certain local region of the input neurons (this certain local region is named receptive field of the neuron), which makes CNN different from ordinary neural networks (made of fully-connected layers). This characteristic solves the obstacle of

ordinary neural networks that they could hardly handle the full images as model's inputs since the necessary learnable parameters (weights) will be too large to compute.

Another important characteristic of CNN is parameter sharing, which means that a specific feature map (consists of a 3D-dimension activations) is obtained from dot product computation with one single fixed filter (or kernel). And this also dramatically decrease the learnable model parameters, leading CNN to be efficient enough to handle computer vision problem.

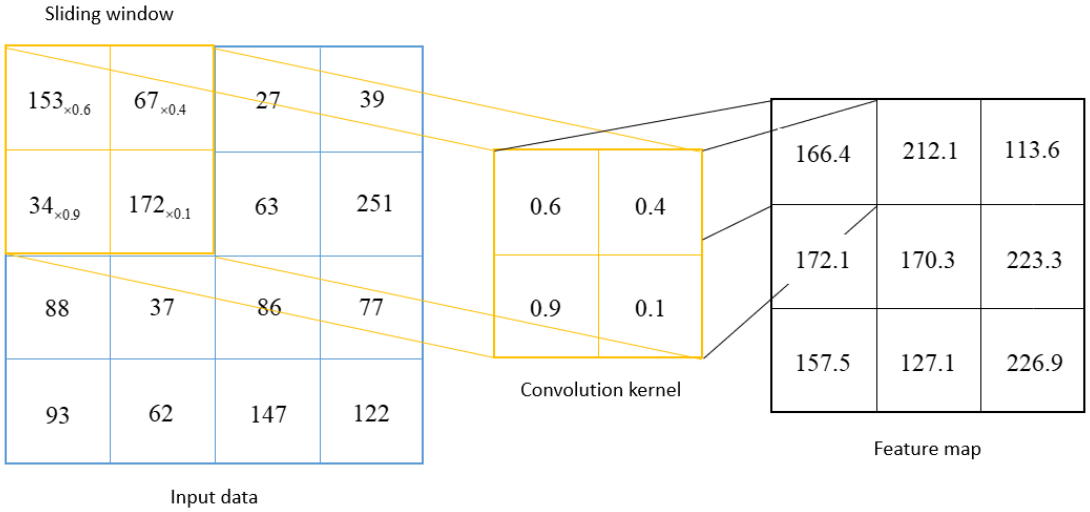


Figure 2. 1. An example of 2-D convolution

Convolution layer is the core component of CNN. Here we give an example of 2-D convolution. The size of convolutional kernel is the receptive field of the neuron in feature map. Every neuron in the feature map is computed by elementwise multiplying the input data (could be an image or the activations from previous layer) with the sequentially sliding convolutional kernel (the stride here is 1).

2.3 Video super-resolution with deep learning

We will then introduce the related technologies of deep learning based video super-resolution. Although it is similar to SISR, there are several notable highlights specifically for VSR task.

2.3.1 Fusion of multiple frames

It is critical for VSR model to leverage the underlying temporal relevance between neighboring frames instead of treating them as a set of independent single images, and this important step is fusion. On the one hand, temporal relevance is potentially informative for

video processing to explore, on the other hand, treating frames as a set of independent images will easily lead to incoherent HR results which is undesirable for the video restoration task.

Most early VSR works such as Kappeler et al. [7] and Tao et al. [8] utilize convolution layers to perform the fusion on multiple LR frames, which could be seen as an automatic feature extraction and integration of the whole frames involved. Recent trend of conducting fusion is to adopt networks with recurrent architecture to gradually fuse multiple frames and gather necessary information by receiving the previous HR output as the input for the next inference, which is much more efficient than previous fusing method since the high frequency details are able to be re-used in such models.

And in this thesis, we will focus on the VSR with recurrent fusion architecture in order to alleviate its shortcomings as introduced in chapter 1.2 and improve the performance.

2.3.2 VSR with recurrent fusion architecture

There are two famous recurrent architecture based VSR models with different types of loss functions (objective functions): FRVSR [10] and TecoGAN [14]. We will briefly introduce them and focus on the technologies related to this thesis.

2.3.2.1 Frame-Recurrent Video Super-Resolution (FRVSR)

FRVSR is the first proposed end-to-end trainable VSR model that adopted the recurrent architecture, it efficiently uses the previously inferred HR estimate to super-resolve the subsequent frame.

It is demonstrated that this frame-recurrent architecture naturally encourages the output frames to be temporally consistent, and the characteristic of re-using the high frequency details increases both the VSR performance and efficiency.

2.3.2.1.1 Framework overview of FRVSR

The framework overview and the losses involved in FRVSR are shown in Figure 2.2 and Figure 2.3 respectively.

A learnable optical flow estimation network (FNet) receives current LR frame I_t^{LR} and previous LR frame I_{t-1}^{LR} as inputs. Then, it generates predicted flow maps F^{LR} so that provide the approximate reference for the subsequent warping the previous HR frame I_{t-1}^{HR} to the current frame, following the procedure in [29]. This is a motion estimation process based on an assumption that the motion changes between neighboring HR frames are similar to the LR

version, which aims to provide more accurate information for inferring the next HR estimate, named alignment in VSR. The distance between warped previous LR frames and the current LR frame I_t^{LR} is then used as loss to train FNet.

After that, through a space-to-depth transformation [22] which extracts shifted low-resolution grids from the input and places them into the channel dimension, the warped previous HR estimate along with the current LR frame are received as the inputs of the reconstruction network (SRNet), generating the estimated HR result I_t^{est} for the current frame. The distance between the estimate I_t^{est} and ground truth HR frame I_t^{HR} is used as loss to train the SRNet. The training process of the two learnable models aim to minimize the following loss functions (in a way of L^2 loss) respectively for FNet and SRNet to optimize the parameters:

$$L_{flow} = \left\| WP(I_{t-1}^{LR}, F^{LR}) - I_t^{LR} \right\|_2^2 \tag{2.1}$$

$$L_{sr} = \| I_t^{est} - I_t^{HR} \|_2^2 \tag{2.2}$$

After completing the training, this end-to-end model can directly infer VSR based on given LR input frames.

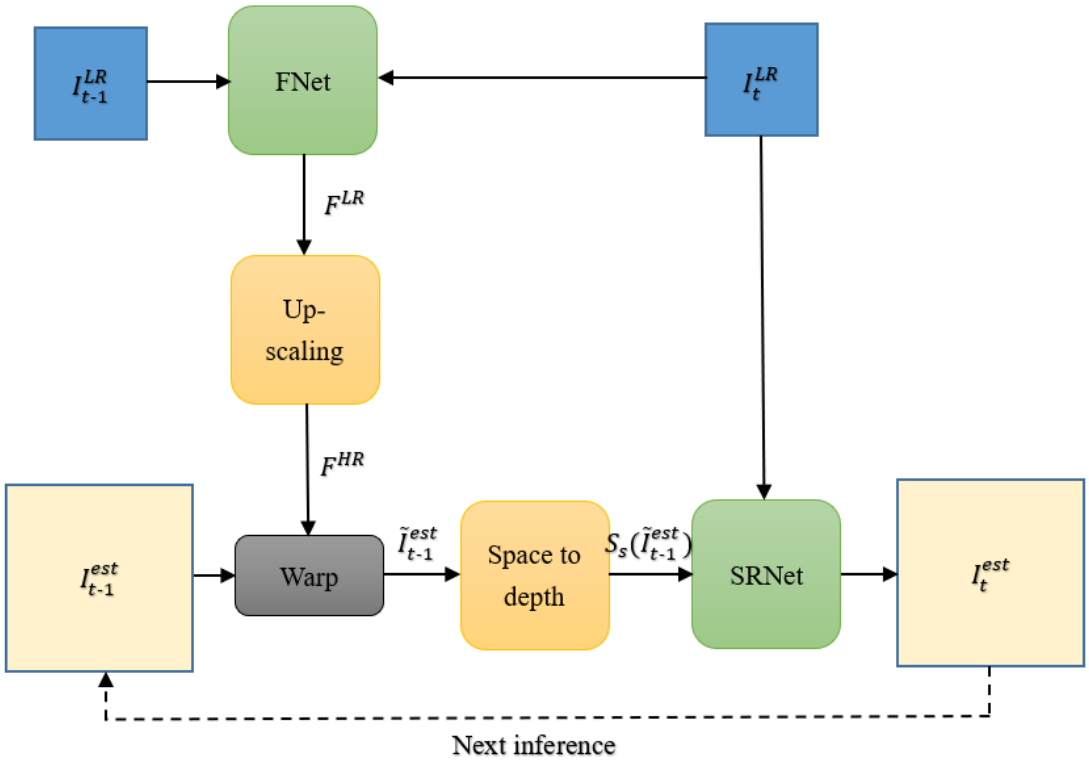


Figure 2. 2. Framework overview of the FRVSR

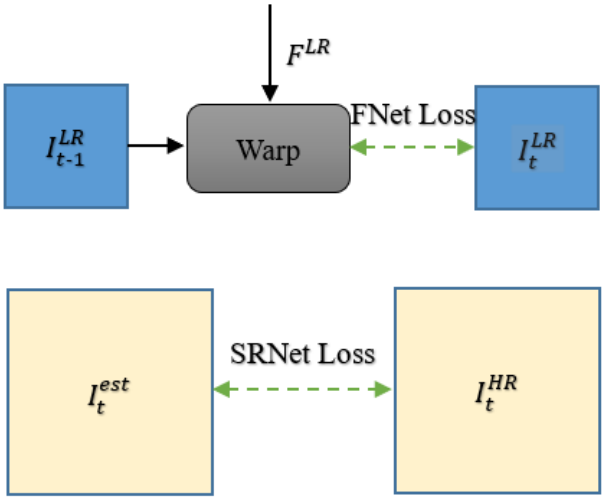


Figure 2. 3. Losses in FRVSR

2.3.2.1.2 Network details and related technologies of SRNet and FNet

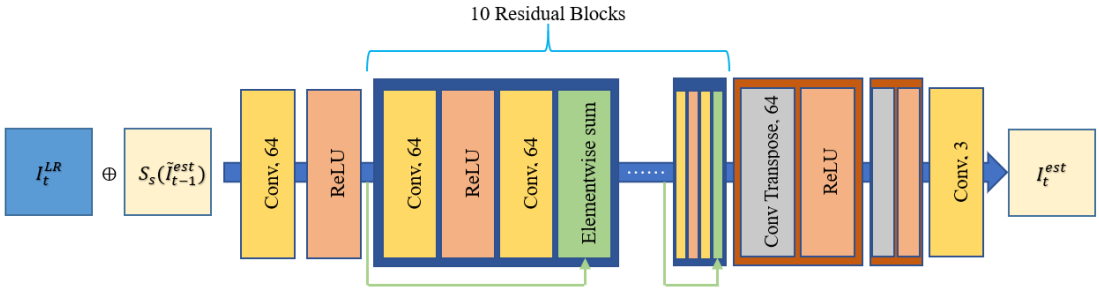


Figure 2. 4. The network architecture of SRNet

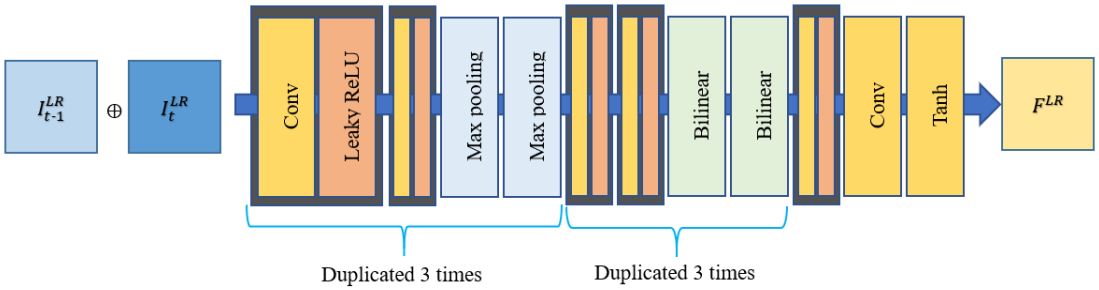


Figure 2. 5. The network architecture of FNet

Figure 2.4 and Figure 2.5 illustrate the network architectures of SRNet and FNet, in which ⊕ indicates the concatenation of the inputs in the channel dimension (depth), the number followed with Conv indicates the number of convolution kernels (filters), which also determined the number of output channels of the layer. Both networks are based on CNN, and they are trained jointly in FRVSR.

There are two notable techniques that deserve to be discussed except for CNN in SRNet: one is transposed convolution layer, and the other is local residual learning.

Transposed convolution layer is also known as deconvolution layer, it is also one of the upsampling techniques like interpolation-based methods. It learns to upsample the input feature maps to upscaled images in an end-to-end manner by performing an opposite version of convolution. Specifically, it predicts and upscales the targeted image resolution by zero-padding and performing convolution. Transposed convolution layer differs from the interpolation-based methods in that it would adaptively (after training) introduce extra information other than only manipulating the inputs' own signals.

According to the analysis of ResNet [15], with the depth of the whole network increases, a learning degradation problem will occur which impede the training. And this issue could be largely alleviated by introducing some shortcuts between layers and optionally learning the residuals between the final targeted output and the input (appropriate for image translation tasks such as SR). The local residual learning is to locally add several shortcuts between the middle layers of the deep neural networks, which benefits the learning procedure by improving the learning efficiency.

On the other hand, FNet simply followed an encoder-decoder style architecture based on CNN. It is not necessary to follow exactly the same structures they adopted since they took the balance between result quality and model complexity when constructing these architectures. It is free to substitute any specific networks with similar functions for them. In fact, recently there are many methods with more complex optical flow estimation such as [16], [17], [18], pre-trained neural networks based perceptual loss functions such as [26], [27], and GAN discriminators such as [23], [24], [28] for substitution.

2.3.2.2 TecoGAN and Ping-Pong (PP) loss

In TecoGAN, the authors intentionally keep the generator part the same with FRVSR to demonstrate the benefits brought by their proposed spatio-temporal discriminator module, which is based on Generative Adversarial Networks [30] (GANs, could be seen as learnable loss functions that get joint training with generator and supervise the generated outputs), aiming at generating perceptually realistic HR outputs.

We would though focus on another main contribution they proposed: Ping-Pong (PP) loss. The authors also noticed one of the drawbacks brought by the recurrent architecture mentioned in chapter 1.2: the accumulating noisy artifacts frame by frame. And this issue is also discovered in a variety of recurrent architectures. Their solution is to introduce a bi-directional loss function to supervise the long-term consistency, the overview of the PP loss is shown below in Figure 2.6.

It is necessary to first duplicate the input frame sequence in order to make it become a symmetric Ping-Pong sequence (where Ping refers to the forward pass and Pong refers to the backward pass). By doing this, a symmetric Ping-Pong sequence of the HR output from the frame-recurrent generator (FRVSR) could be obtained. Similar concept is used in robotic control algorithms [25]. Finally, as removing the noisy artifacts accumulated along with frames is desirable, as well as the output results should be perfectly symmetric, they trained the networks with those extended PP sequences and constrain the generated outputs to be symmetric by introducing a loss function during the training:

$$L_{pp} = \sum_{t=1}^{n-1} \| I_t^{HR} - I_t^{HR\prime} \|_2^2 \tag{2.3}$$

Although the PP loss successfully removes the easily accumulated artifacts in recurrent architecture based VSR, it spent twice training and inferring cost directly because the inference data needs to be doubled.

And most importantly, the solution they considered is to directly suppress this effect, instead of exploring the fundamental cause behind this problem induced by the recurrent architecture. Consequently, the PP loss could only solve one of the shortcomings brought by recurrent architecture at the cost of doubled training and inference expense, other limitations, such as rapid losing high frequency details that are desired to be preserved longer, still remain to be solved.

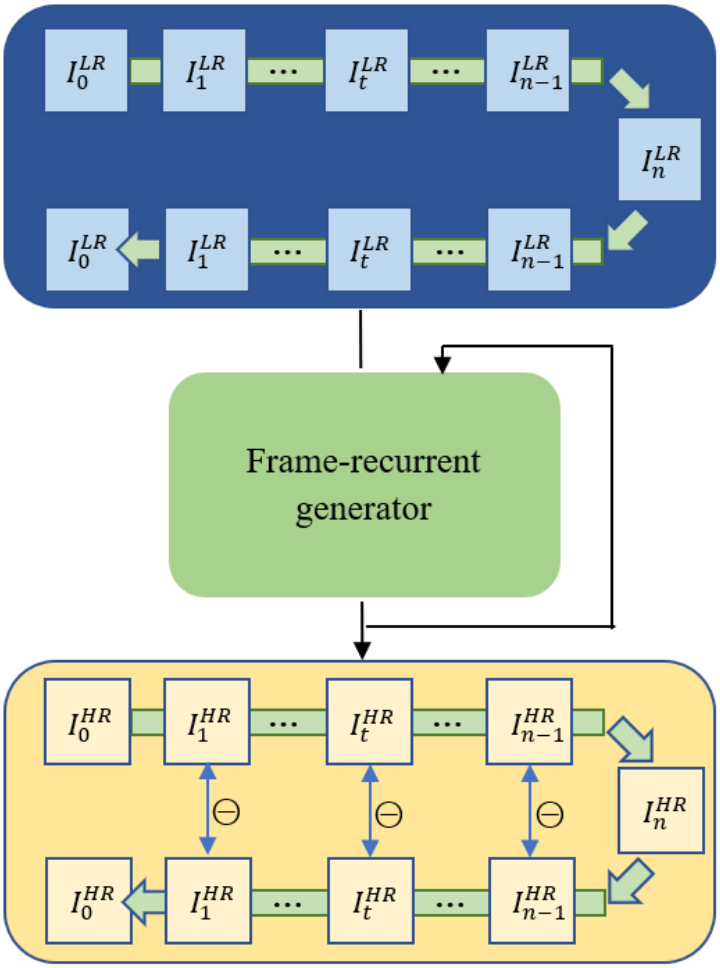


Figure 2. 6. The overview of PP loss proposed in TecoGAN

The VSR models with different fusion schemes introduced above are explicitly fusing the frames without considering spatial informativeness that adaptive to different locations and frames. In this thesis, we would take the spatial informativeness into account and propose selective fusion for solving the problem brought by the recurrent architecture in VSR. Since we choose to fix the problem through analyzing the fundamental cause, it is feasible for us to largely alleviate all the shortcomings it would induce at the same time. Besides, we will demonstrate the extra computation cost that proposed method produce is very limited compared to the solution of PP loss in TecoGAN.

2.4 Quality assessment of SR

Quality assessment in SR means to evaluate the objective visual attributes or subjective perceptual feedbacks of estimated HR images or videos. Although different assessment methods could be inconsistent to each other, they would respectively represent some specific aspects of the generated outputs. Since the subjective assessment based on human perception is inefficient and of unstable accuracy, we choose to evaluate the results with computation based assessing methods.

According to [19], the objective quality assessment methods are mainly divided into three types: full-reference methods which comparing with reference images, reduced-reference methods which comparing with extracted features, and no-reference methods without any reference images. In deep learning based SR tasks, there usually exists LR-HR training pairs and test dataset for easily obtaining the reference data, so performing full-reference methods is appropriate for SR since it is efficient and of computation based accurate assessing results. Next we will introduce the full-reference assessment methods we adopted.

2.4.1 Peak Signal-to-Noise Ratio (PSNR)

Peak signal-to-noise ratio is widely used for quality assessment in image/video restoration tasks which measure the distortion extent of the target image compared to the reference. PSNR is defined by the maximum pixel value and the mean squared error (MSE) between the ground truth image and the corresponding reconstructed image.

Given the ground truth image I with the pixel number of N, and its corresponding reconstructed image $\tilde{I}$, the PSNR between I and $\tilde{I}$ is defined as follows:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{M^2}{\frac{1}{N}\sum_{i=1}^{N}(I(i)-\tilde{I}(i))^2}\right) \tag{2.4}$$

where M equals to the maximum pixel value of the image, for example, when images are using 8-bit representations, M equals to 255 which is the maximum value of pixels. As we can observe that the PSNR only concerns with pixel level difference between the reconstructed image and the ground truth, it is often inconsistent with perceptual assessments. Nevertheless, it still accurately reveals the inherent performance in a fair way that provides metrics for literature comparisons, and there is no absolute perceptual evaluation metric. As the result, PSNR is still the most widely used evaluation metric for SR.

2.4.2 Structural Similarity (SSIM)

SSIM considers the human visual system characteristic of adapting to object structures, it is used to measure the structural similarity through luminance, contrast and structures.

Given the ground truth image I with the pixel number of N, the luminance and contrast are respectively estimated as the mean and standard deviation of the image intensity as follows:

$$\mu_I = \frac{1}{N}\sum_{i=1}^{N} I(i) \tag{2.5}$$

$$\sigma_I = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(I(i) - \mu_I)^2} \tag{2.6}$$

where $I(i)$ indicates the intensity of the i-th pixel of image I, and the SSIM between the ground truth image I and the reconstructed image $\tilde{I}$ could be calculated by:

$$\mathrm{SSIM}(I, \tilde{I}) = \frac{(2\mu_I\mu_{\tilde{I}} + c_1)(2\sigma_{I\tilde{I}} + c_2)}{(\mu_I^2 + \mu_{\tilde{I}}^2 + c_1)(\sigma_I^2 + \sigma_{\tilde{I}}^2 + c_2)} \tag{2.7}$$

where $\sigma_{I\tilde{I}} = \frac{1}{N-1}\sum_{i=1}^{N}(I(i) - \mu_I)(\tilde{I}(i) - \mu_{\tilde{I}})$ is the covariance between I and $\tilde{I}$, c_1 and c_2 are constants for stabilizing.

It is notable that although SSIM is also a full-reference metric like PSNR, but it differs from PSNR in that PSNR estimates the absolute errors between the ground truth and the reconstructed images, and on the other hand, SSIM is based on a perception model which incorporates the prior knowledge of the structural information and perceptual phenomena such as luminance masking and contrast masking, hence it better reflects the perceptual quality and is also widely used in assessing the SR model.

Chapter 3 Proposed Approach

By incorporating the proposed selective fusion into the recurrent architecture based VSR model, there are several architectural improvements need to be applied. We will introduce them in detail and further discuss the framework of selective fusion module and the design principle behind it.

3.1 Framework of proposed method based VSR with recurrent architecture

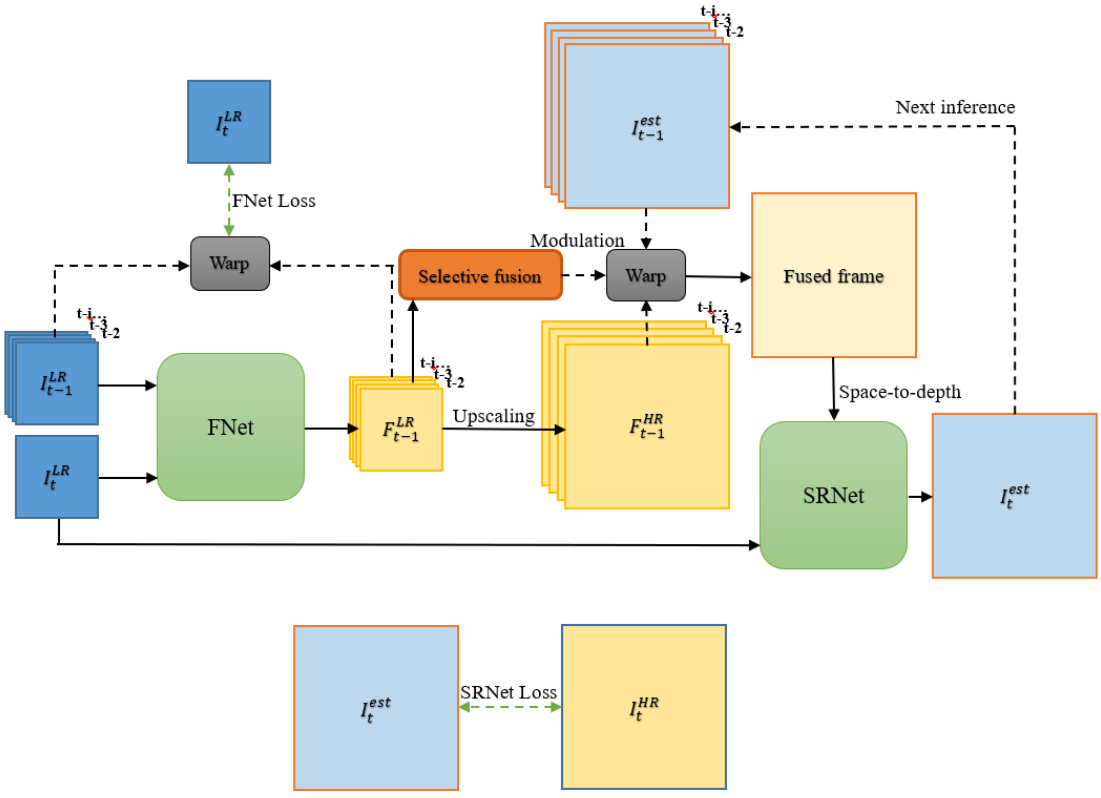


Figure 3. 1. Framework overview of the proposed method

Figure 3.1 illustrates an overview of the video super-solution model with recurrent architecture and selective fusion module, where the FNet and SRNet are still the deep learning based model with learnable parameters. The whole model is still an end-to-end model that conducts joint training. The most notable difference between FRVSR and the proposed method in architecture is that it is necessary for most of the modules in the proposed method to process

multiple frames or features arranged in channel dimension since the model as a whole is no more manipulating only one previous estimated frame. We divide the model into three main stages: motion alignment stage, selective fusion stage and reconstruction stage, and respectively introduce them in detail.

3.1.1 Motion alignment stage

In this stage, our main goal is to accurately align informative details as much as possible that would be potentially beneficial for the reconstruction module. Specifically, we need to warp the motions of previously generated outputs based on an assumption that the motion changes between neighboring frames of HR space are similar to the motion changes between their corresponding LR version. And this sort of motion changes is able to be expressed by the optical flow maps between two neighboring frames.

As the first step, the optical flow maps between LR frame I_t^{LR} and previous LR frames I_{t-1}^{LR}, I_{t-2}^{LR},…I_{t-i}^{LR} in a given video sequence could be estimated by the trained FNet in order, where i equals to the maximum number of previously generated HR frames when inferring the targeted output I_t^{est}. Then normalized flow maps are given by:

$$\{F_{t-1}^{LR}, F_{t-2}^{LR}, \dots, F_{t-i}^{LR} = FNet(I_t^{LR}, I_{t-i}^{LR}) \in [-1,1]^{H\times W\times 2}\} \tag{3.1}$$

where $H \times W \times 2$ denotes the value of three channels (height, weight, and depth) of LR video frame. The generated flow maps represent the predicted movements of each pixels in I_{t-1}^{LR}, I_{t-2}^{LR},…I_{t-i}^{LR} with reference to current LR frame I_t^{LR}.

According to the previously mentioned assumption, we are going to utilize the upscaled LR flow maps to similarly predict the motion changes for warping the HR frames. By applying bilinear interpolation, we are able to efficiently obtain the corresponding HR flow maps:

$$\{F_{t-1}^{HR}, F_{t-2}^{HR}, \dots, F_{t-i}^{HR} = UP(F_{t-i}^{LR}) \in [-1,1]^{sH\times sW\times 2}\} \tag{3.2}$$

in which s denotes the scaling factor which is an inherent variable that decide the upscaling size of SR models ($s = 4$ for this work).

Then the alignment stage could be completed through warping the previously estimated HR frames I_{t-1}^{est}, I_{t-2}^{est},…I_{t-i}^{est} to current frame I_t^{est}:

$$\{\tilde{I}_{t-1}^{est}, \tilde{I}_{t-2}^{est}, \dots, \tilde{I}_{t-i}^{est} = WP(I_{t-i}^{est}, F_{t-i}^{HR})\} \tag{3.3}$$

3.1.2 Selective fusion stage

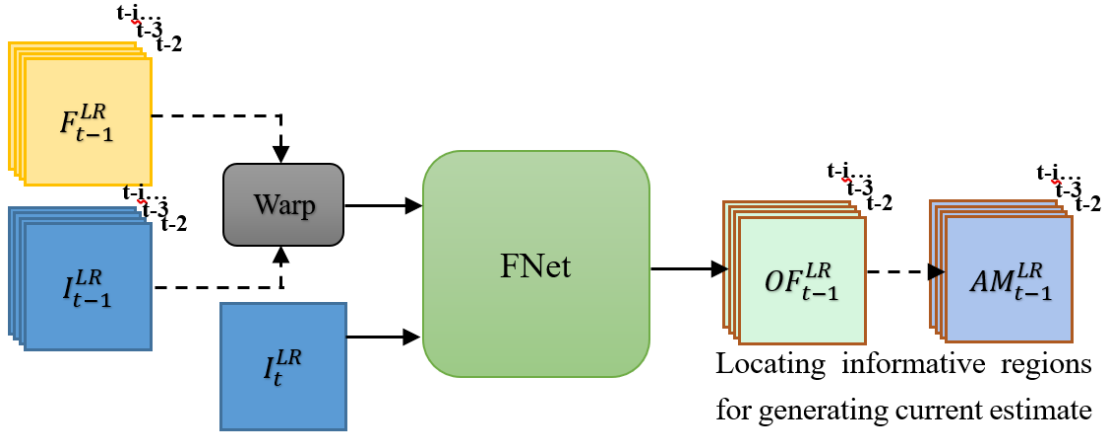


Figure 3. 2. The selective fusion module

The framework of proposed selective fusion module is shown in Figure 3.2. Inspired by the assumption in alignment stage that utilized the universal characteristics between LR and HR sequences to bring underlying beneficial information, we also aimed to further explore the characteristics of LR videos that correlated with the reconstructed HR counterparts.

The module is designed based on an assumption that the warped previously generated HR frames/regions with the more successful warping should have been more informative for reconstruction.

Suppose obstacles that impede the warping process, for example, occlusions and blurs have occurred between two neighboring frames, it would be not possible for the pixels of some certain influenced regions of previous frames to be accurately assigned a position in current frame, which represents those regions are of poor alignment results and less informativeness. If we could exclude these poorly aligned regions through a selective screening performing on all the warped previously estimated HR frames $\tilde{I}_{t-1}^{est}, \tilde{I}_{t-2}^{est}, \ldots, \tilde{I}_{t-i}^{est}$, we are able to gather the most informative regions among them, and fuse them into one single frame used for reconstructing the current HR estimate.

By feeding the warped previous LR frames $\tilde{I}_{t-1}^{LR}$, $\tilde{I}_{t-2}^{LR}$,…$\tilde{I}_{t-i}^{LR}$ (with reference to the current LR frame I_t^{LR}) and the current LR frame I_t^{LR} once again into the FNet, a set of optical flow maps OF_{t-1}^{LR}, OF_{t-2}^{LR},…OF_{t-i}^{LR} could be obtained. Since there exists a ground truth LR current frame for examination, these flow maps are able to reflect the warping (alignment) quality of certain regions of the warped previous LR frames: regions with lower distances of OF^{LR} are expected to be better aligned and thus informative. Then we are able to locate informative regions in the warped previous LR frames $\tilde{I}_{t-1}^{LR}$, $\tilde{I}_{t-2}^{LR}$,…$\tilde{I}_{t-i}^{LR}$ and produce binary attention

maps (a group of matrices) AM_{t-1}^{LR}, AM_{t-2}^{LR}, $\cdots \mathrm{AM}_{t-i}^{LR}$, where "1" indicates the location of minimum optical flow distance.

Next, according to the similarity between LR and HR video sequences, we can instruct the fusing process of multiple warped previously estimated HR frames based on the generated attention maps, which constructs the whole selective fusion stage. The whole stage generates a fused frame that adaptively aggregated the high frequency details that are of more value for reconstructing the current HR frame, it could be expressed as follows:

$$I^{sf} = SF(\sum_{i=1}^{t} \tilde{I}_{t-i}^{est}, \sum_{i=1}^{t} WP(I_{t-i}^{LR}, F_{t-i}^{LR}), I_t^{LR}) \tag{3.4}$$

3.1.3 Reconstruction stage

With a space-to-depth transformation, the fused frame I^{sf} and the current LR frame I_t^{LR} are fed into SRNet, generating the current HR estimate I_t^{est}. The final output of this model can be expressed as follows:

$$I_t^{est} = SRNet(I_t^{LR} \oplus S_s(SF(\sum_{i=1}^{t} \tilde{I}_{t-i}^{est}, \sum_{i=1}^{t} WP(I_{t-i}^{LR}, F_{t-i}^{LR}), I_t^{LR}))) \tag{3.5}$$

3.2 Training objectives

The whole model is still end-to-end trainable. In the training stage, we aim to minimize the following loss functions to optimize the learnable parameters in FNet and SRNet:

$$L_{flow} = \sum_{i=1}^{t} \left\| WP(I_{t-i}^{LR}, F_{t-i}^{LR}) - I_t^{LR} \right\|_2^2 \tag{3.6}$$

$$L_{sr} = \| I_t^{est} - I_t^{HR} \|_2^2 \tag{3.7}$$

3.3 Issue about extra computation cost

It seems that introducing quite a lot additional processing of multiple frames will largely increase the computational complexity. However, in fact, most of the introduced additional procedures such as the bilinear upscaling and warping are performed at very limited computation cost. On the other hand, additional procedures with complex computations such as calculating through FNet and SRNet are all conducted in LR space, which also leads to very limited additional cost. As the consequence, the proposed selective fusion based VSR with recurrent architecture only introduce a little additional computation cost with reference to traditional scheme in FRVSR.

We will demonstrate the detailed comparisons of inference time in the next chapter.

Chapter 4 Experiments and results

We independently train and infer the recurrent architecture based VSR models with three different settings: traditional scheme (FRVSR), with average fusion, and with selective fusion. The superiority of our proposed method would be demonstrated by analyzing both the quantitative and qualitative results.

4.1 Average fusion for comparison

To investigate the effectiveness of the design principle of our proposed selective fusion module, except for comparing the proposed scheme with traditional FRVSR as the reference, we additionally trained another model with the average fusion scheme to exclude other factors that might influence the fairness of the experiments such as more sufficient training dataset for FNet.

In average fusion scheme, the whole model architecture stays unchanged compared to the proposed selective fusion scheme. Multiple previously warped estimates $\tilde{I}_{t-1}^{est}, \tilde{I}_{t-2}^{est}, \ldots, \tilde{I}_{t-i}^{est}$ are also utilized and fused into a single frame. The difference is that those multiple frames are simply performed universal elementwise addition and average with reference to the channel dimension.

4.2 Implementation details

4.2.1 Experimental environments

The recurrent architecture based VSR model with selective fusion is implemented in TensorFlow 1.13. We train and evaluate the models with different settings on a Nvidia GeForce GTX 1080Ti GPU with 11G memory. The experiments are conducted under the OS of Ubuntu 18.04.

4.2.2 Dataset for training and testing

In order to obtain the dataset for training and testing, 250 HR video clips from vimeo.com are collected, in which each clip consists of 120 HR frames. Then the ground truth HR frames can be obtained by down-sampling the collected raw frames by a factor of 2. The corresponding LR frames are produced by applying down-sampling every 4-th pixel for super-resolution scaling factor $s = 4$ and Gaussian blur with a standard deviation $\sigma = 1.5$.

Ground truth HR frames are cropped into training patches of spatial size 128×128 and corresponding LR patches are therefore cropped into 32×32. Batch size is set to be 4, while each sample in the batch is composed of 10 consecutive cropped frame pairs, which means one batch contains 40 cropped frame pairs. Through applying Xavier initialization [20] to the learnable networks and training them utilizing the Adam optimizer [21] with a fixed learning rate of 10^{-4}. The entire training process consists of 200k batches.

Another 10 HR-LR video clips which also collected from vimeo.com are used for the testing dataset of quantitative evaluation.

4.3 Experiments and results analysis

4.3.1 Quantitative evaluation and analysis

Table 4. 1. PSNR evaluation (dB)

	Traditional scheme (FRVSR)	**VSR with average fusion**	**VSR with selective fusion**
Video clip 1	26.6363	26.5836	**26.8030**
Video clip 2	26.4544	26.4990	**26.7456**
Video clip 3	28.4751	29.1327	**29.7333**
Video clip 4	29.6128	29.3145	**29.8711**
Video clip 5	25.1272	25.0542	**25.1501**
Video clip 6	24.7346	24.7317	**25.0341**
Video clip 7	25.7143	25.6234	**25.7337**
Video clip 8	25.8559	25.8690	**25.8638**
Video clip 9	26.4753	26.3891	**26.7131**
Video clip 10	26.7029	26.4923	**26.7364**
Average PSNR	26.5789	26.5690	**26.8384**

Table 4. 2. SSIM evaluation

	Traditional scheme (FRVSR)	**VSR with average fusion**	**VSR with selective fusion**
Video clip 1	0.8773	0.8739	**0.8883**
Video clip 2	0.8806	0.8845	**0.8878**
Video clip 3	0.8550	0.8508	**0.8797**
Video clip 4	0.8650	0.8722	**0.8835**
Video clip 5	0.8563	0.8509	**0.8686**
Video clip 6	0.8431	0.8460	**0.8537**
Video clip 7	0.8704	0.8769	**0.8779**
Video clip 8	0.8661	0.8583	**0.8699**
Video clip 9	0.8792	0.8756	**0.8891**
Video clip 10	0.8894	0.8808	**0.8897**
Average SSIM	0.8682	0.8670	**0.8788**

We inferred the trained models with different three settings and respectively evaluated PSNR and SSIM of their outputs on different testing video clips, the results are respectively shown in Table 4.1 and Table 4.2.

As we can see in both tables, our proposed recurrent architecture based VSR model with selective fusion outperforms the other two settings in every testing samples. The average quality of inferred videos has increased by 0.2595 dB in PSNR and 0.0106 in SSIM compared to traditional setting (FRVSR). The results can demonstrate the effectiveness of the design principle of our proposed selective fusion in the capability of efficiently fusing information. And on the other hand, the model with average fusion fails to compete the traditional scheme in any index, since it is possible that distant frame will introduce unnecessary and even wrong information aggregating for generating current estimate, and lead to degraded performance.

Table 4. 3. Inference time evaluation

	Traditional scheme (FRVSR)	**VSR with average fusion**	**VSR with selective fusion**
Average inference time (ms/frame)	**153.9**	155.7	155.9

The evaluation results of inference time of different settings are shown in Table 4.3. As we supposed in chapter 3, proposed method is proved to be efficient enough that it only brings a little additional computation cost compared to the traditional scheme. Without introducing any extra parameters of the learnable models, the increased computation of FNet is efficiently conducted in LR space and can even further enhance the training process in a way similar to performing data-augmentation (frames that originally distant from I_t^{LR} are also received as training samples), which implicitly upgrades the robustness of FNet.

4.3.2 Qualitative evaluation and analysis

Qualitative evaluation is conducted through inferring the models of different settings with a 120-frame video clip from a license open movie “Tears of Steel” [31] that allowed to be demonstrated. Table 4.4 shows the quantitative evaluation results on this particular video clip for different settings.

Table 4. 4. Quantitative results for “Tears of Steel”

	Traditional scheme (FRVSR)	**VSR with average fusion**	**VSR with selective fusion**
PSNR (dB)	28.0922	27.0896	**35.0070**
SSIM	0.8993	0.8876	**0.9639**

The results of our proposed method largely surpass other settings in both PSNR and SSIM. It is possibly because a severe effect of gradually enhanced noisy artifacts occurs in the traditional scheme for this particular video clip as we could observe in the last frame of the generated outputs with different settings shown in Figure 4.1.

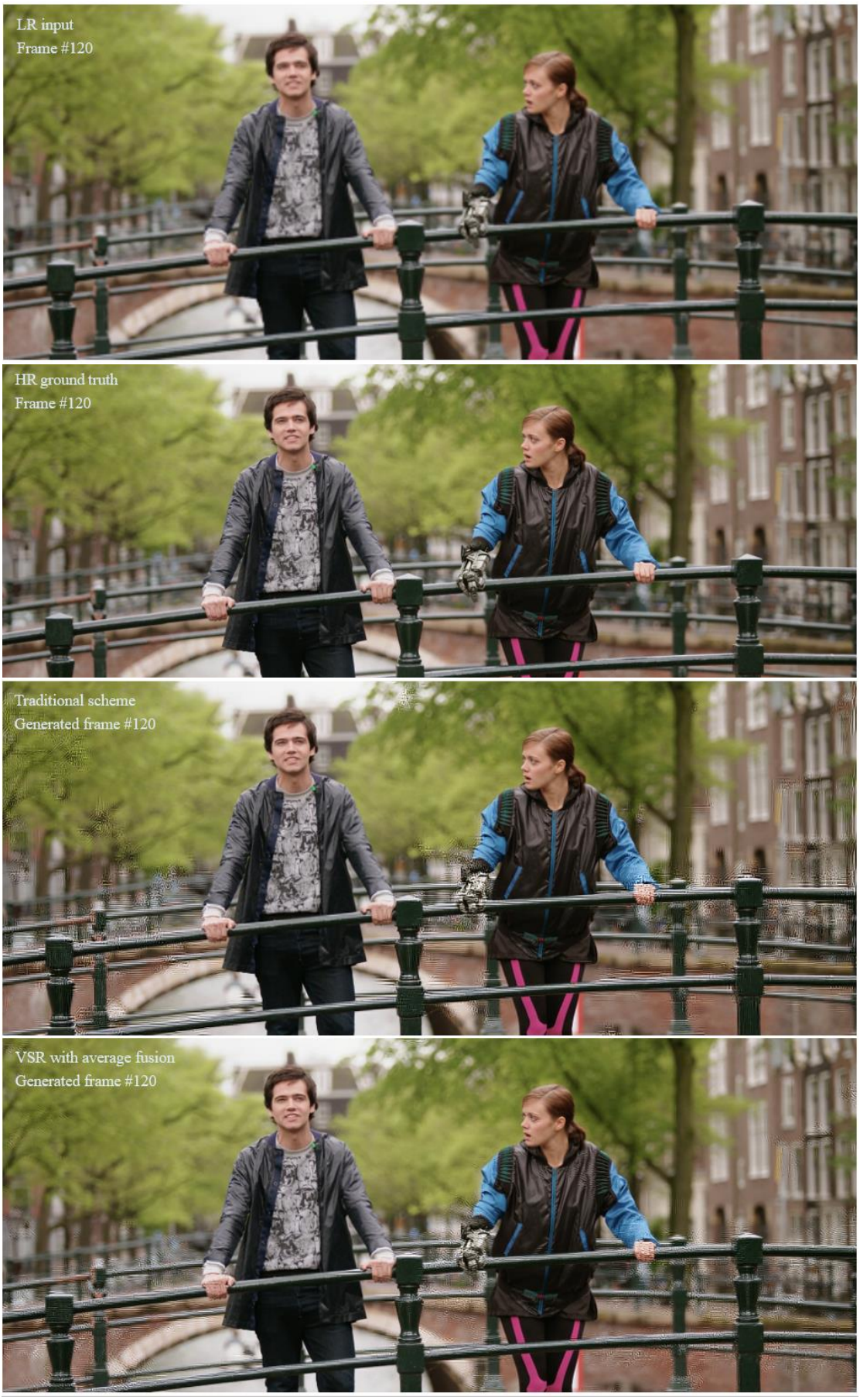

LR input
Frame #120
HR ground truth
Frame #120
Traditional scheme
Generated frame #120
VSR with average fusion
Generated frame #120

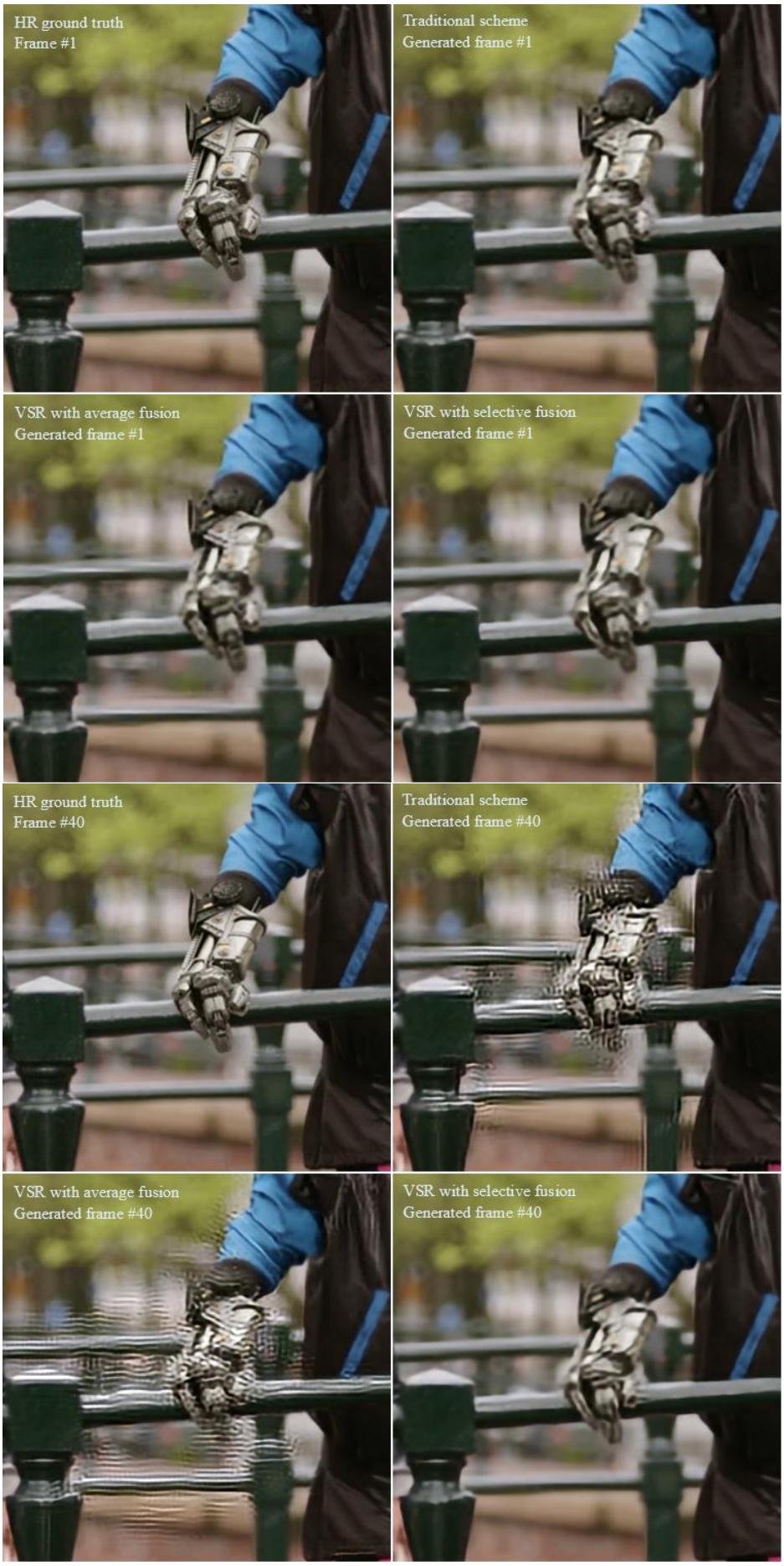
Figure 4. 1. Comparison of the last frame of the generated outputs with different settings (picture source: [31])

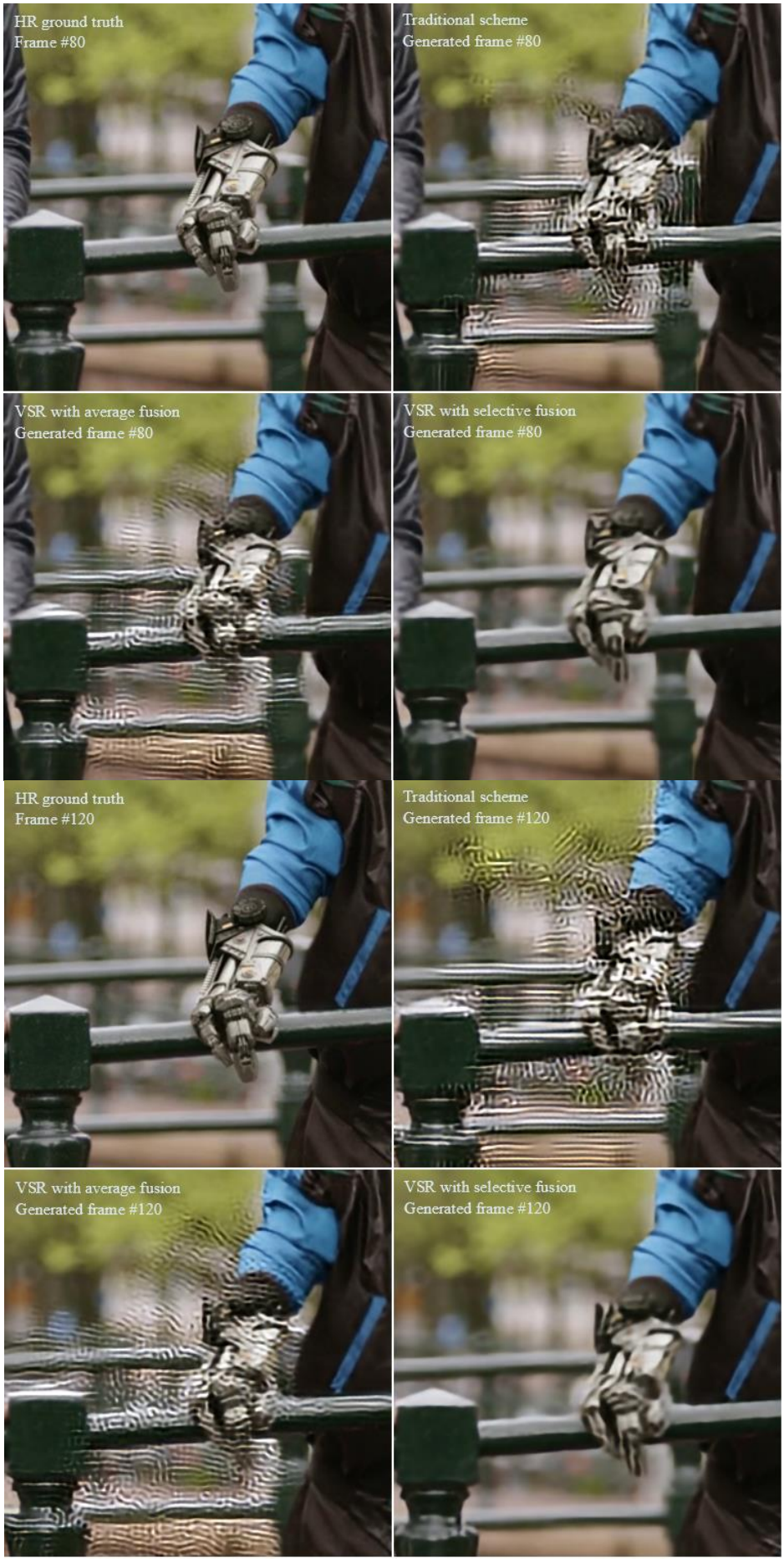
The last generated frame of traditional scheme aggregates the most severe artifacts with the largest area. The model tries to sharpen the details to lower the loss while being incapable of retrieving better input information or discarding undesirable noise, since it over relies on one single fixed previously generated estimate during a long recurrent process, which seriously limits the performance.

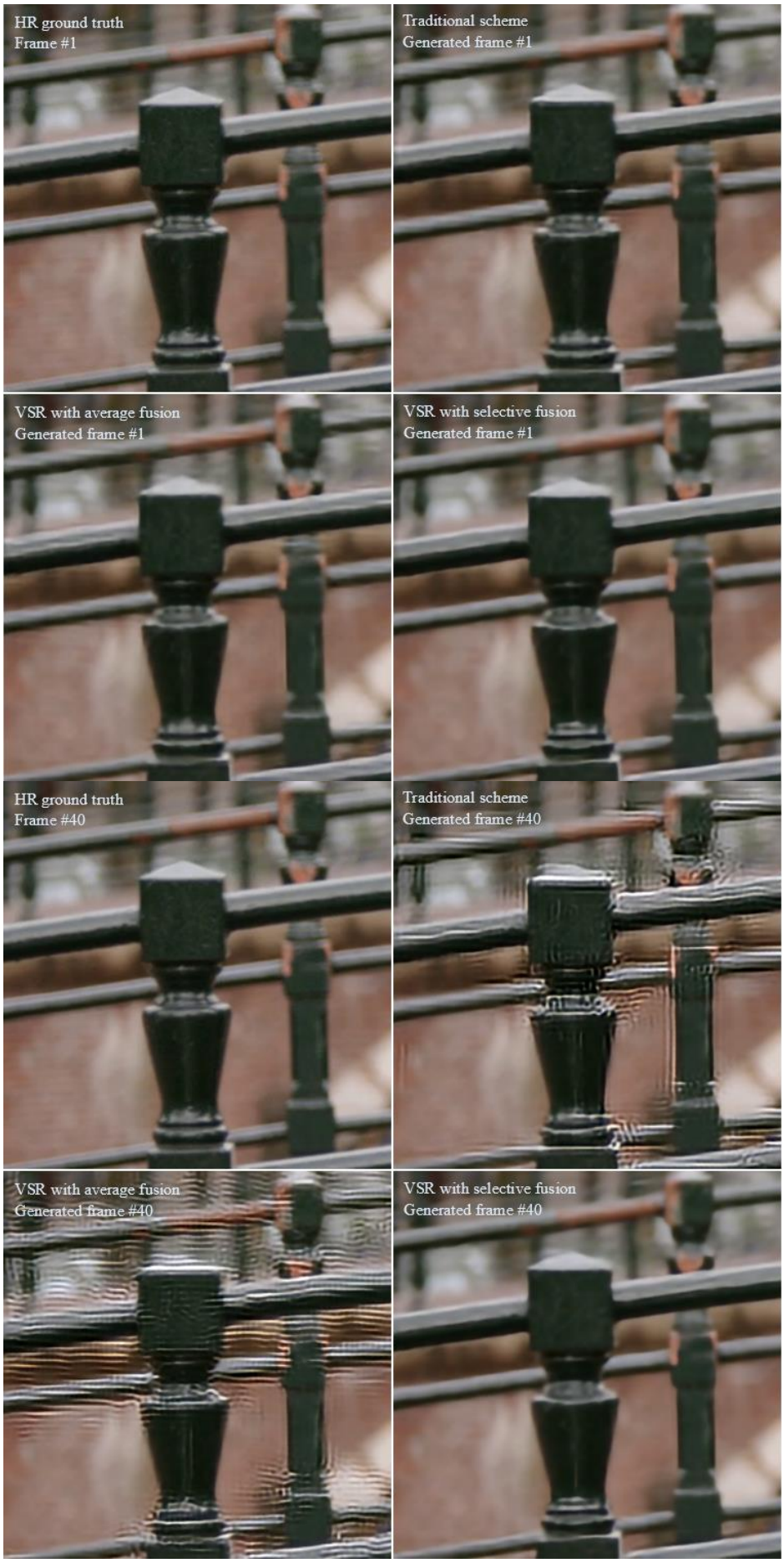
The recurrent architecture based VSR model with average fusion slightly alleviates this undesirable effect simply since it averages the poorly informative details with other better informative regions, hence it still produces those sequentially strengthened noisy artifacts.

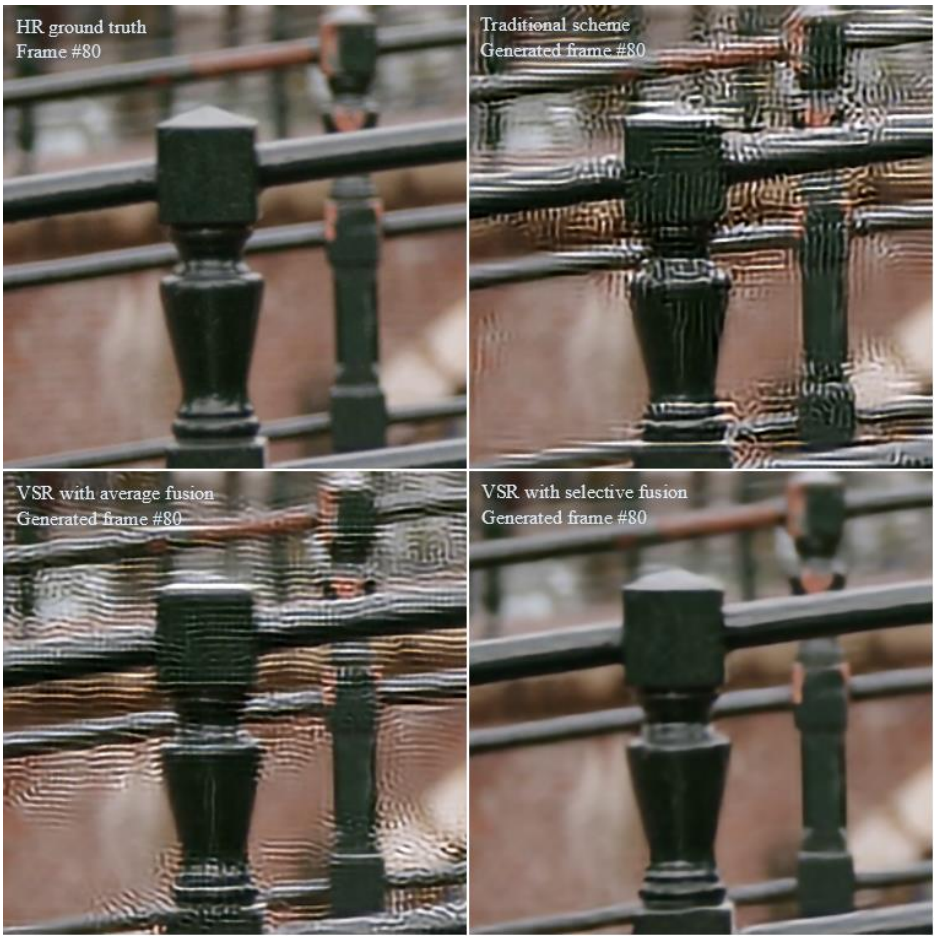
The proposed model with selective fusion correctly and efficiently removes these undesirable stains and gives a clean estimated HR sequence with satisfying SR quality, it reflects that our proposed method is able to effectively retrieve and fuse the information in need among the previously generated estimates that are beneficial for inferring the next HR frame. It is also the proof that our designed fusion principle, which concerning the fundamental cause of the shortcomings brought by recurrent architecture in VSR models, is effective.

More results of cropped patches are shown in Figure 4.2. It is illustrated that the gradually enhanced noisy artifacts in traditional scheme have been perfectly removed in proposed method, which clearly demonstrate the superiority of our proposed method in visual performance.


HR ground truth
Frame #1
Traditional scheme
Generated frame #1
VSR with average fusion
Generated frame #1
VSR with selective fusion
Generated frame #1
HR ground truth
Frame #40
Traditional scheme
Generated frame #40
VSR with average fusion
Generated frame #40
VSR with selective fusion
Generated frame #40


HR ground truth
Frame #80
Traditional scheme
Generated frame #80
VSR with average fusion
Generated frame #80
VSR with selective fusion
Generated frame #80
HR ground truth
Frame #120
Traditional scheme
Generated frame #120
VSR with average fusion
Generated frame #120
VSR with selective fusion
Generated frame #120


HR ground truth
Frame #1
Traditional scheme
Generated frame #1
VSR with average fusion
Generated frame #1
VSR with selective fusion
Generated frame #1
HR ground truth
Frame #40
Traditional scheme
Generated frame #40
VSR with average fusion
Generated frame #40
VSR with selective fusion
Generated frame #40


HR ground truth
Frame #80
Traditional scheme
Generated frame #80
VSR with average fusion
Generated frame #80
VSR with selective fusion
Generated frame #80

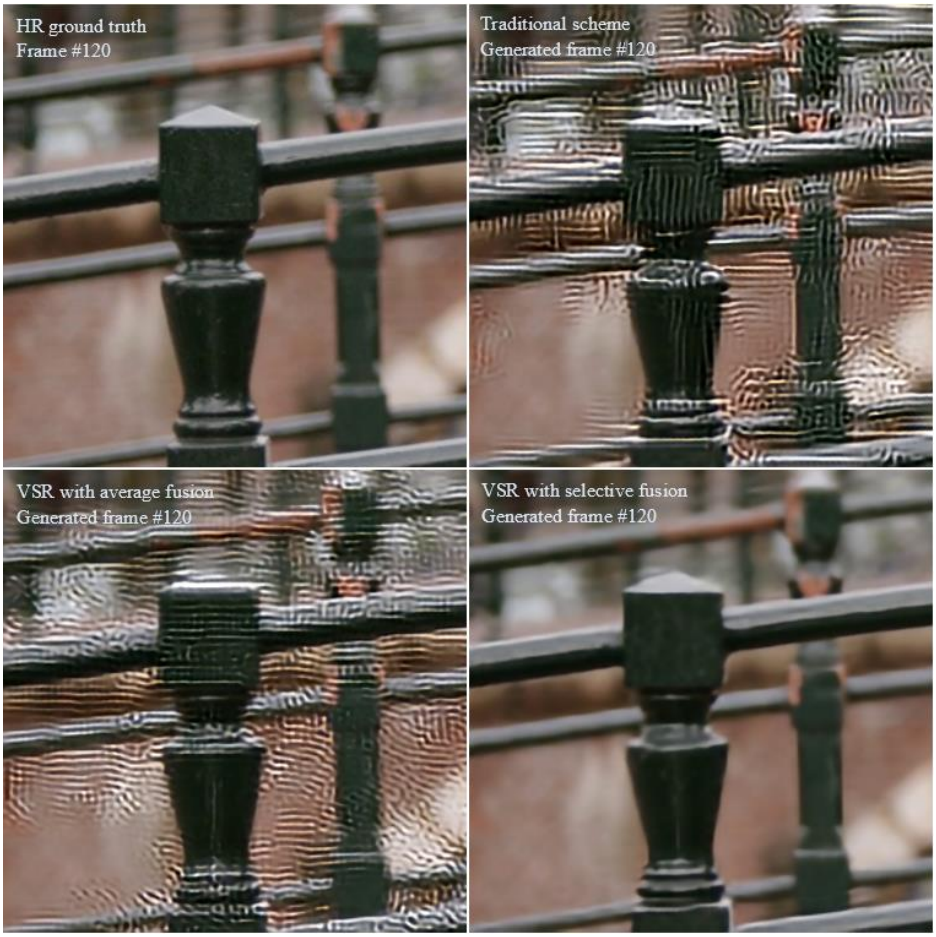


Figure 4. 2. Detailed inference results of cropped patches (picture source: [31])

Chapter 5 Conclusion

We proposed selective fusion for addressing the followed shortcomings induced by recruiting recurrent architecture in video super-resolution model such as the rapid information loss and sequentially strengthened noisy artifacts which largely limits the performance of generated outputs.

Based on the designing assumption that the warped previously generated HR frames/regions with the more successful warping should have been more informative for reconstruction, and without changing the scale and dimensions of the inputs of Super-Resolution Network (SRNet), the selective fusion module efficiently and successfully gathers and fuses informative details from previous generated HR estimates according to their informativeness, instead of simply relying on one single previous generated HR frame. The proposed method is demonstrated to be able to improve video super-resolution performance while introducing neglectable additional computation cost.

Chapter 6 Appendix

6.1 List of academic achievements

International conference:

Z. Gong, T. Hori, H. Watanabe, T. Ikai, T. Chujoh, E. Sasaki, and N. Ito: “A Selective Fusion Module for Video Super Resolution with Recurrent Architecture,” International Workshop on Advanced Image Technology, IWAIT 2020, No.43, Jan. 2020

Domestic conference:

Z. Gong and H. Watanabe : “An Evaluation of The Impact of Dataset Bias in Pretrained VGG Network on The Performance of Neural Network Based Style Transfer,” IEICE General Conference, BS-4-19, Mar. 2019

Bibliography

[1] R. Keys, "Cubic convolution interpolation for digital image processing," IEEE Transactions on Acoustics, Speech, and Signal Processing,

vol. 29, no. 6, pp. 1153–1160, 1981.

[2] C. E. Duchon, "Lanczos filtering in one and two dimensions," Journal of Applied Meteorology, vol. 18, no. 8, pp. 1016–1022, 1979.

[3] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient

profile prior," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2008.

[4] A. Marquina and S. J. Osher, "Image super-resolution by TVregularization and Bregman iteration," Journal of Scientific Computing, vol. 37, no. 3, pp. 367–382, 2008.

[5] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based superresolution," IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 56–65, 2002.

[6] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3791–3799, 2015.

[7] A. Kappeler, S. Yoo, Q. Dai & A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," IEEE Transactions on Computational Imaging 2(2), 109-122, 2016.

[8] X. Tao, H. Gao, R. Liao, J. Wang, & J. Jia, "Detail-revealing deep video super-resolution," Proc. ICCV pp. 4472-4480, 2017.

[9] Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., & Huang, T., "Robust video super-resolution with learned temporal dynamics," Proc. ICCV pp. 2507-2515, 2017.

[10] M. S. Sajjadi, R. Vemulapalli, & M. Brown, "Frame-recurrent video super-resolution," Proc. ICCV pp. 6626-6634, 2018.

[11] M. Haris, G. Shakhnarovich, & N. Ukita, "Recurrent Back-Projection Network for Video Super-Resolution," Proc. CVPR pp. 3897-3906, 2019.

[12] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A. K. Katsaggelos, "Softcuts: a soft edge smoothness prior for color image super-resolution," IEEE Transactions on Image Processing, vol. 18, no. 5, pp. 969–981, 2009.

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in Proceedings of the European Conference on Computer Vision, pp. 184–199, 2014.

[14] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, & N. Thuerey, "Learning temporal coherence via self-supervision for GAN-based video generation," arXiv preprint arXiv:1811.09393, 2018.

[15] K. He, X.Zhang, S. Ren, & J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox., "Flownet: Learning optical flow with convolutional networks," In Proceedings of the IEEE international conference on computer vision, pp. 2758-2766, 2015.

[17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462-2470, 2017.

[18] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4161-4170, 2017.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, 2004.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249-256, 2010.

[21] D. Kingma and J. Ba. Adam, "A method for stochastic optimization," In ICLR, 2015.

[22] W. Shi, J. Caballero, F. Husz´ar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1874-1883, 2016.

[23] C. Ledig, L. Theis, F. Husz´ar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681-4690, 2017.

[24] M. S. M. Sajjadi, B. Sch¨olkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," In Proceedings of the IEEE International Conference on Computer Vision, pp. 4491-4500, 2017.

[25] Suraj Nair, Mohammad Babaeizadeh, Chelsea Finn, Sergey Levine, and Vikash Kumar, "Time reversal as self-supervision. arXiv preprint arXiv:1810.01128, 2018.

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," In European Conference on Computer Vision. Springer, pp. 694–711, 2016.

[27] Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao, "Perceptual adversarial networks for image-to-image transformation," IEEE Transactions on Image Processing 27, 8, 4066–4079, 2018.

[28] You Xie, Erik Franz, Mengyu Chu, and Nils Thuerey, "tempoGAN: A Temporally Coherent, Volumetric GAN for Super-resolution Fluid Flow," ACM Transactions on Graphics (TOG), 37(4), 1-15, 2018.

[29] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," In IEEE Transactions on Computational Imaging, 2(2), 109-122, 2016.

[30] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y., "Generative adversarial nets," In Advances in neural information processing systems, pp. 2672-2680, 2014.

[31] (CC) Blender Foundation | mango.blender.org, "Tears of steel," <https://mango.blender.org/>, 2011.