Improved EDVR Model for Robust and Efficient Video Super-Resolution

Yulin Huang and Junying Chen*
School of Software Engineering, South China University of Technology
Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
Guangzhou, China
815018345@qq.com, jychense@scut.edu.cn

Abstract

Computer vision technologies are increasingly commonly used in daily life, and video super-resolution is gradually drawing more attention in the computer vision community. In this work, we propose an improved EDVR model to tackle the robustness and efficiency problems of the original EDVR model in video super-resolution. First, to handle the blurring situations and emphasize the effective features, we devise a preprocessing module consisting of rigid convolution sub-modules and feature enhancement sub-modules, which are flexible and effective. Second, we devise a temporal 3D convolutional fusion module, which can extract information in image frames more accurately and rapidly. Third, to better utilize the information in feature maps, we design a new reconstruction block by introducing a new channel attention approach. Moreover, we use multiple programmatic methods to accelerate the model training and inference process, making the model useful for practical applications.

1. Introduction

1.1. Background

Nowadays, computer vision technology plays an important role in research and industry communities, but high-resolution information is usually not easy to obtain, especially for videos. Hence, video super-resolution is a good solution. However, video super-resolution algorithms are facing two challenges. On the one hand, the accuracy is not so satisfactory. On the other hand, many video applications require high-speed models, even real-time models. Traditional algorithms, such as bicubic interpolation and bilinear interpolation, cannot gain the ideal output, while machine learning-based algorithms can get better results than previous methods, but usually at the cost of time-consuming training and enormous model parameters.

In the past few years, super-resolution has made remarkable improvements thanks to deep learning. Early studies [3, 20] regard video as a temporal sequence of images, so they cannot exploit the temporal redundancy among neighboring image frames sufficiently. Recent studies [2, 24] represent videos as a set of movements and solve the super-resolution problem with a more elaborated pipeline which usually consists of four parts, namely feature extraction, frame alignment, fusion, and reconstruction. To obtain high-quality output, the video super-resolution model should be able to: (1) align neighboring frames accurately and effectively fuse them, and (2) reconstruct output information by effectively combining the fused features. Moreover, the model should be accelerated to fit practical video super-resolution applications.

1.2. Related Works

Since the pioneering work of SRCNN [5], deep learning-based methods are playing an increasingly important role in super-resolution solutions. Super-resolution in a single image has been extensively studied. Since image super-resolution is a feature extraction and 2D generation problem, deep learning models demonstrate great accuracy and robustness. As for videos, more redundant information is hidden in the temporal space, so temporal alignment is one of the most significant procedures. Deep learning-based solutions show remarkable advantages when handling large motions. Another difficulty is to get information in aligned frames accurately. VESPCN [2] uses convolutional architecture to fusion features in neighboring frames to overcome such difficulty. Moreover, D3DNet [25] uses deformable convolution and 3D convolution in video super-resolution, which successfully lift the result to a higher level. In general, video super-resolution greatly benefits from deep learning techniques.

*Corresponding author.
This work was supported in part by the National Natural Science Foundation of China under Grant 61802130, and in part by the Guangdong Natural Science Foundation under Grant 2019A1515012152 and Grant 2021A1515012651.

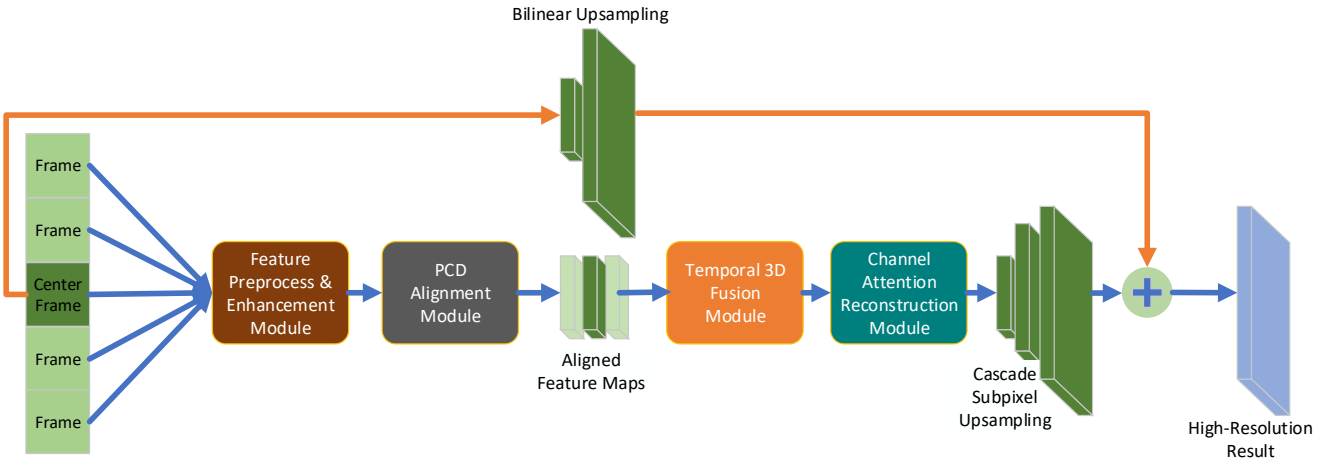

Figure 1. The architecture of the proposed model. This model consists of a feature preprocessing and enhancement module, a PCD alignment module [22], a temporal 3D convolution module, and a channel attention reconstruction module.

Feature extraction: Feature extraction is the first step of the model. Most existing models use a simple convolutional layer to perform feature extraction [2], but such a method usually produces bad results in blurred and distorted situations. In the enhanced deformable convolutional network for video restoration (EDVR) [22], the authors apply a self-ensemble strategy to solve the problem, but such a large model reduces the computational efficiency. The way to balance accuracy and computational efficiency is worth investigating.

Frame Alignment: Frame alignment is based on motion estimation. Most existing models explicitly estimate optical flow fields to align neighboring frames [2, 24]. This type of methods can estimate slow motions effectively but has worse results for quick motions. Another way to align frames is to estimate frames using convolutions [10, 21], which can enhance the robustness of the model.

Fusion: Fusion is a significant component of the video super-resolution model. Earlier video super-resolution models usually adopt convolutions or recurrent architectures to fuse aligned frames. Recent models use temporal attention mechanism [22] to assign different weights to aligned frames because frames are not equally informative to the reconstruction. Such a method greatly improves fusion results but makes the model too complex and inflexible. A good way to fuse aligned frames will bring significant improvement.

Computational Efficiency: Computational efficiency is an important factor that limits the applications of the model. Recent models based on deep learning gain good results at the cost of huge computational time and GPU memory consumption so that they are not ideal solutions. Auto mixed precision (AMP) [18] is a novel solution, which uses scaling and resizing to store 32-bit floating-point numbers to a smaller type so as to save both computational time and GPU memory consumption. Another good way to accelerate the deep learning model is applying the just-in-time (JIT) compilation technique. Under the PyTorch framework, dynamic code can be represented using type-specific TorchScript representation [4], and the route tracing technique is used to reorganize layers to make JIT more effective.

1.3. Our Contributions

Based on the EDVR model [22], we design an innovative model, which is more robust and efficient. Compared to the EDVR model, this work concentrates more on video super-resolution tasks. The contributions of this work are listed in the following.

First, we introduce a novel feature preprocessing and enhancement module, which (1) uses multi-scale extraction to process early feature extraction result in order to improve robustness, (2) uses cascade well-designed rigid convolution block to improve the robustness and training stability, and (3) uses feature enhancement block to improve the results in some bad but regular situations (*e.g.*, too dark, distorted, etc.).

Secondly, we introduce a temporal 3D fusion module to fuse information in neighboring frames. Compared to EDVR's temporal and spatial attention (TSA) module [22], our module uses 3D convolution to extract attention information more robustly. Furthermore, our model reduces 10% parameters using the weight-sharing technique.

Thirdly, we design a channel attention reconstruction module by adopting the theory in SENet [7], so as to replace the cascade residual blocks in EDVR. As a result, our model can use the information in different channels more selectively and accurately.

Finally, our model also focuses on execution efficiency. We apply the mixed-precision training [18], JIT compila-

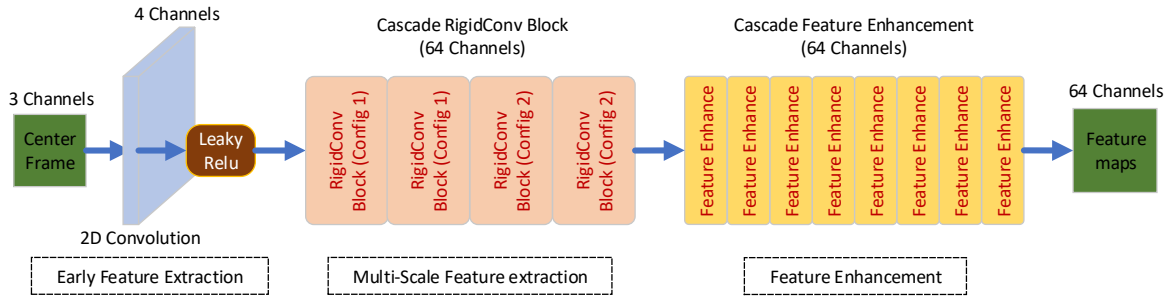

Figure 2. The structure of the feature preprocessing and enhancement module.

tion, ATEN library, new CUDA programming approach, and cosine annealing with restart strategy [8] to the model. All of these measures can accelerate the model without losing accuracy.

2. The Proposed Model

2.1. Overview

The overall framework of the proposed model is shown in Fig. 1. The model takes $(2N + 1)$ low-resolution frames as inputs and generates a high-resolution estimation of the center frame (which is also called the reference frame). First, the low-resolution frames are processed by the feature preprocessing and enhancement module to generate feature maps. Second, these feature maps are passed to the PCD Alignment module. This module is originally proposed in the EDVR model [22], and we rewrite it in a new approach to support execution acceleration measures as described in Section 2.5. Then aligned feature maps are sent to the temporal 3D fusion module. By using 3D convolution, the module can fuse the feature robustly. The last part is a reconstruction module, which can get a high-resolution estimation from the fused feature map. By using the channel attention mechanism properly, the reconstruction module performs well.

In addition, we adopt Charbonnier loss [12] as the loss function in the proposed model. This new type of loss function is a good replacement of L1 loss function. Its math representation is $F(x) = \sqrt{x^2 + \epsilon^2}$. This function avoids vanishing gradient phenomenon when $x = 0$, and makes the training process more stable.

2.2. Feature Preprocessing and Enhancement

The feature preprocessing and enhancement module is a combination of feature extraction and optimization. The module consists of two key components: rigid convolution (RigidConv) group and feature enhancement group. The structure of the whole module is shown in Fig. 2. Many models use simple cascade convolutional groups to generate feature maps, which is too rough to get useful feature maps. So we propose a feature preprocessing and enhancement module which solves this weakness and improves the results, generating better feature maps to improve the performance of the whole model.

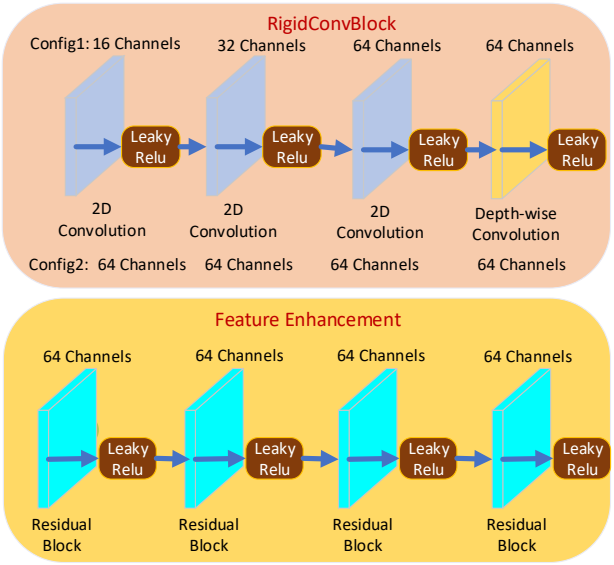

Figure 3. Top: The structure of Rigid Convolution (RigidConv) Block; Bottom: The structure of Feature Enhancement Block.

Aimed at generating feature maps more accurately, inspired by Inception architecture [13], Rigid Convolution Block is proposed. It contains a depth-wise convolutional layer to improve the robustness of the model. Also, we adopt a multi-scale strategy, where the Rigid Convolution uses multi-layers to gradually enlarge the feature map, which can make the model faster and more robust. The structure is shown in Fig. 3 (Top). Like the structure shown in the figure, we set up two different configurations for the block, each has an independent group of parameters. "Config 1" enables the block to deepen the feature maps step by step, while "Config 2" keeps the depth of feature maps unchanged. When cascading RigidConv Block, RigidConv Blocks of "Config 1" is first used to get the feature maps of target channels, and then a proper number of RigidConv Blocks of "Config 2" is used to conduct operations such as deblurring and denoising. Compared to the cascading

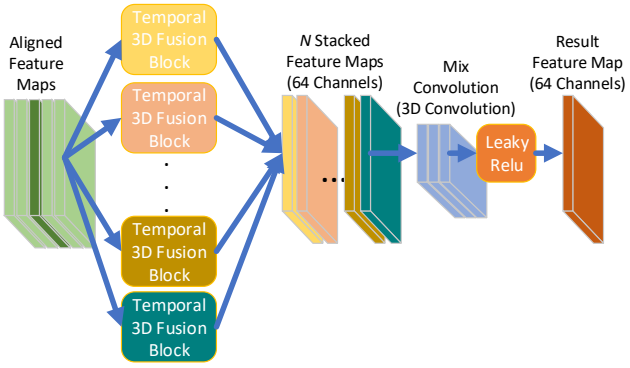

Figure 4. The structure of temporal 3D fusion module.

convolutional feature extraction groups in EDVR, we use Leaky Relu rather than Relu to avoid vanishing gradient problems, and we use a smaller convolution core and more layers rather than large convolution cores to avoid information loss and extract more accurate features.

Feature enhancement is another vital sub-module to optimize the feature map. This sub-module is designed based on inspiration from ResNet and SRResnet. We use the structure similar to the basic block in ResNet [6] to construct the feature enhancement sub-module, as interpreted in Fig. 3 (Bottom). Like SRResNet [14], Batch Norm layers are removed to keep the absolute difference in feature maps. We also adopt Leaky Relu activation to avoid losing gradients in some blocks. The block enhances the feature maps from the feature extraction stage and uses the residual structure to fine-tune the feature maps. Moreover, it is possible to cascade this sub-module to get a better result, so this structure is very flexible.

2.3. Temporal 3D Fusion

The temporal 3D fusion is a novel fusion module using 3D convolution. The most important structure of this module is temporal 3D convolution. The idea is to regard the consecutive frames as a multi-channel feature map and reorganize the frames through the temporal sequence to form the depth of the feature map to extract both temporal and spatial redundancy. The structure of the temporal 3D fusion module is shown in Fig. 4.

The Temporal 3D Fusion Block is the core functional block in the temporal 3D fusion module. The block uses 3D convolution to extract temporal and spatial attention weights rather than using independent weight maps, which is more robust. In the original EDVR[22] model, a TSA model is used for fusion frames by applying attention mechanisms in both temporal and spatial areas. However, the limitation of 2D convolution makes this module only assign one temporal attention value on each frame, and the spatial attention value in each frame is completely independent. 3D convolution can greatly improve this weakness.

Moreover, the block adopts a weight-sharing strategy to reduce the number of parameters because it is common that the motion in the front part is the same as the motion in the rear part in consecutive frames. Inside the block, the feature map of the reference frame appears three times because it is the majority of the result. A depth-wise 3D convolutional layer is used to combine the center, the front, and the rear. The structure is shown in Fig. 5. Compared to using only one convolution core, the weight-sharing technique can also make it easier to adopt other accelerate methods like AMP and parallel computing.

Using the temporal 3D fusion module, we improve model performance and significantly reduce the number of parameters. Taking a temporal 3D fusion module with three blocks as an example, it saves about 10% parameters than the TSA fusion module in EDVR [22].

2.4. Channel Attention Reconstruction

Traditional models, like SRResNet [14], usually give all the channels in feature maps completely the same attention, which is very simple to realize while usually getting good enough results. Therefore, many super-resolution models such as EDVR [22] use them to perform the reconstruction. However, the reconstruction problem is ill-conditioned, so this simple way will greatly limit reconstruction accuracy. Inspired by SENet [7] architecture, we design a new attention mechanism on channel dimension. Like the structure of SRResNet [14], the proposed reconstruction module is also based on residual block. The proposed channel attention is calculated from the maximum value and average value of each channel and uses linear layer and activation layer to calculate. The new basic block is called Reconstruction Block, whose structure is shown in Fig. 6.

In this Reconstruction Block, we design an innovative channel attention sub-module. Like SENet [7], we use a two-stage activation strategy. In stage one, we first calculate the average value and maximum of each channel, then we use the linear layer to squeeze the result into a smaller feature map and use Leaky Relu to eliminate the linear dependency of attention on each channel. In stage two, we use the linear layer to restore the feature map's size and use the Sigmoid function to map the attention values to (0,1) to avoid gradient explosion and eliminate linear dependency. This sub-module is the core component of Reconstruction Block, and the experiment results prove the effectiveness of the combination of maximum and average values.

2.5. Computational Efficiency Improvements

The improvement in computational efficiency is also concerned in this work. We use four major programmatic measures to accelerate the model. First, auto-mixed- precision(AMP) [18] technology is added to the model, and an innovative CUDA compiling technique is used to sup-

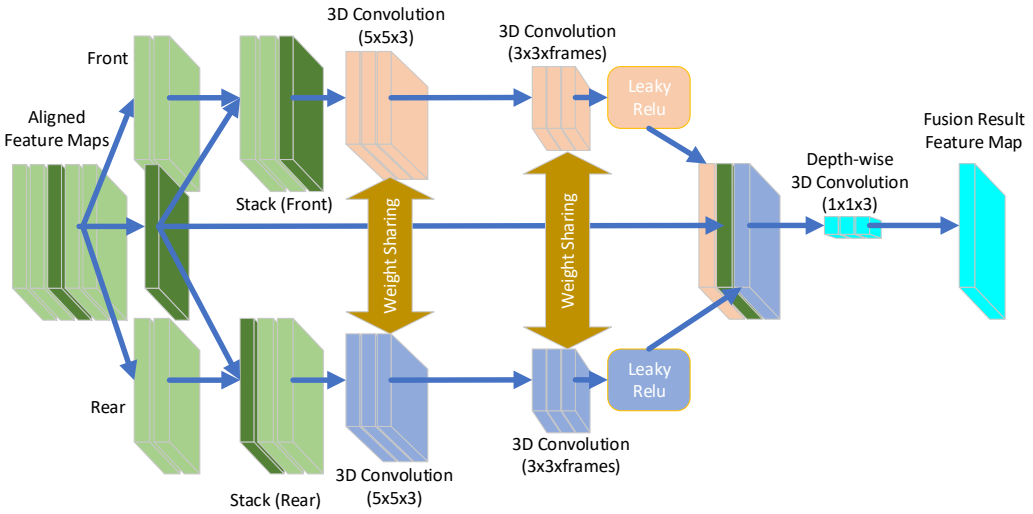

Figure 5. Temporal 3D Fusion Block with weight-sharing.

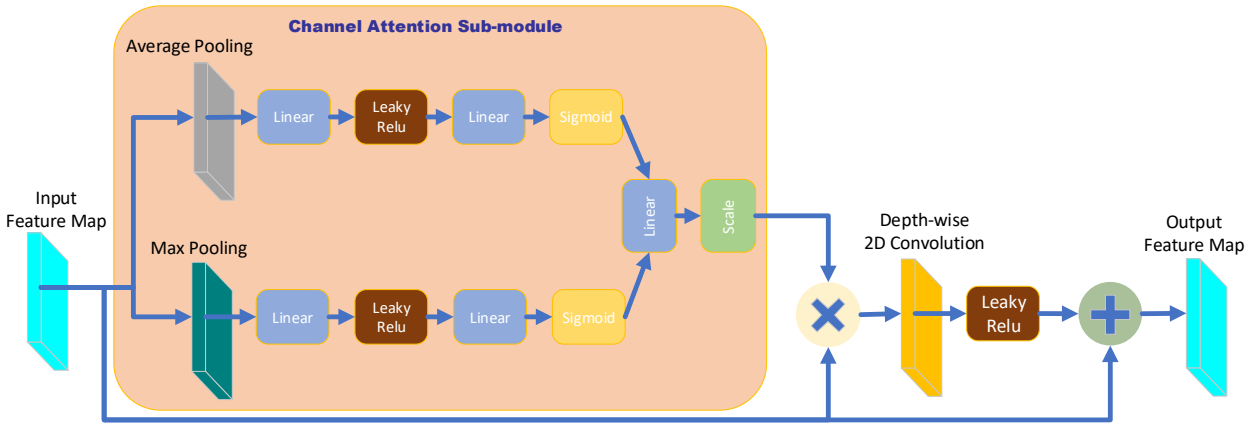

Figure 6. The structure of Reconstruction Block with an innovative channel attention sub-module.

port this technology. This novel CUDA compiling technique uses "ATEN" and "C10" libraries to replace legacy floating-point identifiers, and we use this technique to reconstruct the PCD alignment module [22]. Besides, we use Ninja compiler to enable the PCD alignment module to benefit from the AMP technique. The Ninja compiler can select a proper type for CUDA extension modules to generate suitable linked libraries. Moreover, we applied the latest just-in-time (JIT) compilation technique, including converting and tracing. These techniques enable the module to precompile in whether simple or complex situations.

In addition, we use cosine annealing learning rate scheduler [8] to replace the fixed learning rate. This strategy is proved to be effective and can improve training accuracy and reduce time consumption. When applying this method, we discover an effective way to make it work better. We also apply an automatic restart strategy, which automatically resets the learning rate scheduler to the initial state after processing the whole training dataset. This approach demonstrates its improvement in computational performance.

3. Experiment Results and Discussions

3.1. Experimental Details

In the experiments, we set up a experiment environment with CUDA, and its information is shown in Table 1. In the training stage, we set the depth of feature maps (after feature extraction module) to 64, and use 10 Reconstruction Blocks to form the reconstruction module. We use Adam optimizer [11] to train the model with $\beta_1 = 0.9, \beta_2 = 0.99$, and set the initial learning rate as 5×10^{-5}.

The realistic and dynamic scenes (REDS) dataset [17] and Vimeo90k [26] / Vid4 [15] dataset are used for the experiments. The super-resolution results will be compared qualitatively and quantitatively.

Figure 7. The qualitative super-resolution results.

Item	Value
CPU	AMD Ryzen 3600x
GPU	NVIDIA Geforce RTX2060 Super
Memory	32G,dual channels
Operating System	Microsoft Windows 10,64 bits
Running Environment	Python 3.9.2, PyTorch 1.8.1

Table 1. Experiment Environment Information

Method	PSNR	SSIM
Bicubic	26.2355	0.7319
MFCNN	27.6014	0.7857
RCAN* [27]	28.78	0.8200
TOFlow* [24]	27.98	0.7990
DUF* [9]	28.63	0.8251
EDVR (M)	28.7416	0.8245
EDVR (L)* [22]	31.09	0.8800
Ours	29.7801	0.8552

Table 2. Quantitative Experiment Results on REDS Dataset. '*' indicates that the result is taken from the experiment in EDVR paper [22].

3.2. Super-Resolution Results on REDS Dataset

The REDS dataset contains 30000 frames with 720×1280 resolution, and the scenes in the dataset are diverse and complex. In the experiments, we compare our model with Bicubic Interpolation, MFCNN [1], and EDVR (M, 64 feat version) [22] models in the qualitative study. The qualitative experiment results are shown in Fig. 7. In Fig. 7, the large picture is the ground truth, and we crop the representative area of the super-resolution results of each model to carry on the comparative analysis. As seen from Fig. 7, our model exhibits better super-resolution results, making the representative area clearer and sharper.

On the other hand, in the quantitative experiments, each experiment repeats 4 times, and the mean result is calculated as the final result. The overall final result is shown in Table 2. The evaluation metrics include peak signal to noise ratio (PSNR) and structural similarity (SSIM) [23]. The larger the PSNR and SSIM values, the better the video super-resolution results. The quantitative results demonstrate that our model has better performance than most of the compared models and achieves nearly 4% SSIM improvement when compared to the baseline model EDVR (M). When compared to EDVR (L) which has 128 channels, although our model performs slightly worse, but the number of parameters in our model is 16.8% of that in EDVR (L) (The parameter comparison is shown in Table 3). Table 3 also shows that our model has a similar number of parameters compared to EDVR (M). In summary, our model greatly reduces the timing cost and memory consumption while also has extraordinary performance.

Model	Parameter Count
EDVR (M, 64 channels)	3300131
EDVR (L, 128 channels)	20633827
Ours	3468295

Table 3. Model Parameter Comparison

3.3. Super-Resolution Results on Vimeo90k/Vid4 Dataset

We also conduct quantitative experiments on the Vimeo90k/Vid4 dataset. Vimeo90k is a large dataset based on video clips, so it is suitable to be used as a training dataset. Vid4 is a commonly known test dataset, which is very widely used in video-related tasks. This experiment uses Vimeo90k dataset to train our model and test the model in Vid4 dataset, and we will compare the PSNR, SSIM and VMAF [19] metrics with widely-used models. VMAF is a machine learning-based video quality metric, which can evaluate video quality much closer to human vision. The results are shown in Table 4. Because Vimeo90k and Vid4 datasets only include limited motions, which enable optical-

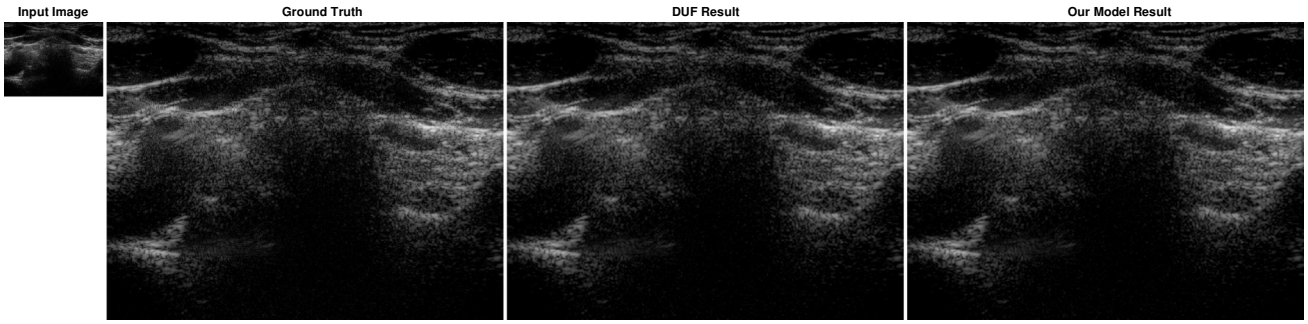

Figure 8. The qualitative results on ultrasonic dataset. From left to right: input, ground truth, DUF model result, and our model result.

Method	PSNR	SSIM	VMAF
Bicubic	22.37	0.6098	45.45
RCAN* [27]	24.02	0.7192	-
MFCNN	22.98	0.6358	36.01
TOFlow* [24]	24.41	0.7428	-
DUF* [9]	25.79	0.8136	-
EDVR (M)	25.38	0.7795	77.08
EDVR (L)* [22]	25.83	0.8077	-
Ours	25.61	0.7956	83.39

Table 4. Quantitative Experiment Results on Vid4 Dataset. '*' indicates that the result is taken from the experiment in EDVR paper [22].

flow-based algorithms and deep models to get good results without sufficient robustness. So in the results, our model performs slightly better than EDVR (M) and worse than DUF [9] and EDVR (L) [22] in PSNR and SSIM. However, in the VMAF metric, Our model shows superior performance than the original EDVR(M), which can show the satisfactory performance of our model.

3.4. Application on Ultrasonic Images and Videos

Super-resolution is a widely discussed topic in processing medical images and videos, helping clinical doctors to make more accurate diagnoses. Nowadays, ultrasonic imaging is a fundamental technology to assist clinicians in diagnoses and treatments. However, ultrasonic devices cannot get high-resolution images or videos due to physical limitations while ensuring deep enough penetration. There are some previous research works about applying image super-resolution technology to improve the quality of ultrasonic images [28, 16]. However, these works ignore the information hidden in the motions. In our research, we use our model to retrieve information behind frames in ultrasonic videos and use continuous motion information to get better super-resolution results than previous image-based algorithms.

We set up experiments on the ultrasonic medical dataset. The quantitative results are shown in Table 5. The results

Model	PSNR	SSIM
Bicubic	28.6992	0.7587
DUF	30.3025	0.7795
Ours	30.3337	0.7820

Table 5. Super-Resolution Quantitative Results on Ultrasonic Dataset

	Whole Model	Without Preprocessing and Enhancement
PSNR	29.7801	28.8824
SSIM	0.8552	0.8197

Table 6. Ablation Study Results of Preprocessing and Enhancement Module

show that the video super-resolution technique can still get a better result than traditional interpolation measures. Because of the limited motion, these videos seldom contain very complex motion. As a result, both DUF and our model can obtain satisfactory outputs, but our model performs slightly better. The qualitative results are shown in Fig. 8, and we can see that the result of our model is very satisfactory.

3.5. Ablation Studies

We conduct ablation studies to validate the effectiveness of each module in our model. All experiments in ablation studies are performed on REDS dataset.

3.5.1 Preprocessing and Enhancement Module

In our model, the preprocessing and enhancement module can eliminate bad features in feature maps and emphasize important features, in order to get more accurate features. We set up an ablation experiment by replacing this module with an identity module. Table 6 depicts the results of this ablation study.

	3 Temporal 3D Fusion Blocks	1 Temporal 3D Fusion Block	TSA Fusion
PSNR	29.7801	29.5439	29.3191
SSIM	0.8552	0.8508	0.8427

Table 7. Ablation Study Result of Temporal 3D Fusion Module

	Our Module	10 Residual Blocks	15 Residual Blocks
PSNR	29.7801	28.2620	28.6276
SSIM	0.8552	0.8065	0.8192

Table 8. Ablation Study Results of Channel Attention Reconstruction Module

3.5.2 Temporal 3D Fusion Module

In this experiment, we compare the results of the proposed temporal 3D fusion module and the TSA fusion module from EDVR [22]. Meanwhile, we also compare the results with different numbers of Temporal 3D Fusion Blocks. The results are shown in Table 7. It is observed that the temporal 3D fusion module produces better results than the TSA fusion module. This is due to the superior performance of 3D convolution and our strategy of parameters sharing, which enable us to use more blocks. Furthermore, the results also show that using more blocks can improve the results to some extent.

3.5.3 Channel Attention Reconstruction Module

In our model, the channel attention mechanism assigns a floating-type weight value to each channel and multiplies it with the original value in the corresponding channel in the feature map, which makes the reconstruction more selective and accurate. In this ablation experiment, to prove the effectiveness of channel attention in reconstruction, we use the reconstruction module in EDVR [22] with two different configurations (10 Residual Blocks and 15 Residual Blocks) for comparison because they have the same depth and similar amount of parameters as compared with our module, respectively. We construct our reconstruction module by cascading 10 Channel Attention Reconstruction Blocks. The comparison results are shown in Table 8.

The results indicate that Channel Attention Reconstruction Block has significant advantages, which obtains better performance than the module consisting of residual blocks with similar or fewer parameters.

3.6. Computing Performance Optimization

To accelerate our model to make it more useful and practical, optimization techniques have been applied to our model. Firstly, the AMP technique is used to enlarge the batch size, such that in our experiment, the batch size is increased from 4 to 6. Secondly, the JIT compilation and route tracing also boost the efficiency of our model. In our experiments, we compare our model with the original EDVR (M) model and our model with different acceleration techniques. The result is shown in Table 9. Each case is examined three times, and the average value is calculated as the final result. The results are shown in Table 9, we can see the superb speed of our model. Moreover, by applying the complete optimization technique, the time cost of each frame is reduced by 29.3%, which is a remarkable optimization.

Configuration	Frame Rate
Original EDVR (M)	2.703 frames/sec (batch size=4)
Our model without optimization	3.030 frames/sec (batch size=4)
Our model with AMP only	3.722 frames/sec (batch size=6)
Our model with all optimizations	4.286 frames/sec (batch size=6)

Table 9. Computing Performance Comparison on REDS Test Cases with Maximum Batch Size

The results indicate that using performance optimization measures can greatly accelerate the model. As video super-resolution is always used in a time-limited situation, our performance optimization measures are meaningful in practical applications.

4. Conclusion

We propose an innovative model based on the EDVR model for robust and efficient video super-resolution tasks. In order to improve the super-resolution performance of the original EDVR model, we focus on feature extraction, frame fusion, and reconstruction improvements. We design an innovative two-stage feature preprocessing and enhancement module and use 3D convolutions and channel attention to form the video super-resolution model. Besides, the temporal 3D convolution method is proposed for fusion, which can fuse aligned frames by learned attention values on both temporal and spatial space. In addition, we use a channel attention mechanism like that in the SENet model, calculating attention weights based on each channel's maximum value and mean value. Moreover, we use the latest technique to accelerate the model, obtaining a faster model with better super-resolution performance.

References

[1] K. J. Abraham Sundar and V. Vaithiyanathan. Multi-frame super-resolution using adaptive normalized convolution. *Signal Image and Video Processing*, 11:357–362, 2017.

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2848–2857, 2017.

[3] Q. Dai, S. Yoo, A. Kappeler, and A. K. Katsaggelos. Dictionary-based multiple frame video super-resolution. In *Proceedings of IEEE International Conference on Image Processing*, pages 83–87, 2015.

[4] Zachary DeVito, Jason Ansel, Will Constable, Michael Suo, Ailing Zhang, and Kim Hazelwood. Using python for model inference in deep learning. *arXiv preprint arXiv:2104.00254*, pages 1–14, 2021.

[5] C. Dong, C. C. Loy, K. He, and Tang X. Learning a deep convolutional network for image super-resolution. In *Proceedings of European Conference on Computer Vision*, pages 184–199, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2017.

[8] Loshchilov I and Hutter F. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of International Conference on Learning Representations*, pages 1–16, 2017.

[9] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3224–3232, 2018.

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

[11] Diederik Kingma and Jimmy Ba. Adam: A method forstochastic optimization. In *Proceedings of International Conference on Learning Representations*, pages 1–15, 2015.

[12] W. Lai, J. Huang, N. Ahuja, and M. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5835–5843, 2018.

[13] Eric-Tuan Le, Iasonas Kokkinos, and Niloy J. Mitra. Going deeper with convolutions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 105–114, 2017.

[15] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2014.

[16] Heng Liu, Jianyong Liu, Shudong Hou, Tao Tao, and Jungong Han. Perception consistency ultrasound image super-resolution via self-supervised cyclegan. *Neural Computing and Applications*, 2021.

[17] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1996–2005, 2019.

[18] Micikevicius P, Narang S, Alben J, Diamos G, Elsen E, Garcia D, Ginsburg B, Houston M, Kuchaiev O, Venkatesh G, and Wu H. Mixed precision training. *arXiv preprint arXiv:1710.03740*, pages 1–12, 2018.

[19] Reza Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In *Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–2, 2017.

[20] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Transactions on Image Processing*, 18(9):1958–1975, 2019.

[21] Y. Tian, Y. Zhang, Y. Fu, and C. Xu. TDAN: Temporally-deformable alignment network for video super-resolution. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3357–3366, 2020.

[22] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1954–1963, 2019.

[23] Zhou 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*, 13(4):600–612, 2004.

[24] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019.

[25] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters*, 27:1500–1504, 2020.

[26] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2443–2452, 2018.

[27] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of European Conference on Computer Vision*, pages 294–310, 2018.

[28] Ningning Zhao, Qi Wei, Adrian Basarab, Denis Kouamé, and Jean-Yves Tourneret. Single image super-resolution of medical ultrasound images using a fast algorithm. In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pages 473–476, 2016.