Video Super-Resolution

Shaomin Xu
sx2311@nyu.edu

Jincheng Liang
jl12003@nyu.edu

Yueyu Hu (Mentor)
yh3986@nyu.edu

***Abstract*—Video super-resolution (VSR) methodologies typically encompass more components compared to their image counterparts, necessitating the exploration of additional temporal dimensions. Consequently, these approaches often adopt intricate designs. In this study, we meticulously examine the EDVR, TGA, BasicVSR, and BasicVSR++ models, subjecting them to testing on the UVG dataset. Through a detailed analysis of the performance variations observed in each model on the UVG dataset, we identify distinct spatial and temporal features in different videos contributing to these differences. Subsequently, we introduce IconVSR+ by optimizing the utilization of spatial features, resulting in superior performance on UVG videos that exhibit suboptimal outcomes with referenced models. We also tried to supervise the model in the frequency domain. The proposed new model WavVSR can better capture high-frequency information and achieve better performance in several UVG scenes. The code is avaliable at https://github.com/JinchengLiang/VSR.**

I. Overview

Shaomin's responsibilities include the following subtasks:

- Read papers EDVR [2], BasicVSR [1] and BasicVSR++ [5].
- Execute the code associated with the BasicVSR model to facilitate a practical understanding of its implementation.
- Test BasicVSR and BasicVSR++ models on UVG [4] dataset to observe their performances on this specific dataset.
- Analyze Evaluate the diverse performance of the referenced models on UVG videos.
- Create a novel IconVSR+ model and assess UVG performance using the model trained on REDS4.

Jincheng is responsible for following subtasks:

- Read TGA [3] and BasicVSR [1] paper.
- Test TGA, IconVSR model on UVG [4] dataset.
- Do complexity analysis on the models.
- Introduce the wavelet-based loss to the BasicVSR. According to the experimental results, the new method WavVSR has better results in several scenarios on UVG data.

II. Project Accomplishment

Shaomin accomplished several subtasks, which encompassed: 1) reviewing the BasicVSR [1] and EDVR [2] papers; 2) implementing the BasicVSR model, previously trained on the REDS4 [6] dataset, on the UDM10 [8] dataset, following the guidelines provided in the referenced paper; 3) evaluating the performance of the BasicVSR and BasicVSR++ model, initially trained on the REDS4 dataset, on the UVG [4] dataset; 4) assess the varied performance of the specified models on UVG videos; 5) develop a fresh IconVSR+ model and evaluate UVG using the IconVSR+ model trained on REDS [6] dataset.

Jincheng completed several subtasks, including: 1) reviewing the TGA [1] and the BasicVSR paper; 2) executing the TGA model trained on the Vid4 [9] dataset and the Vimeo-90K-T [7] dataset, as outlined in the referenced paper. 3) assessing the performance of the TGA model on the UVG [4] dataset; 4) Analyse the complexity of all the models, including parameter size and FLOPs. 5) Introduce wavelet-based loss to the BasicVSR and evaluate on the UVG dataset.

A. Literature Review

To handle large motions, **EDVR** [2] introduce a Pyramid, Cascading, and Deformable (PCD) alignment module, aligning frames at the feature level using deformable convolutions in a coarse-to-fine manner. Additionally, the Temporal and Spatial Attention (TSA) fusion module emphasizes crucial features for restoration both temporally and spatially. EDVR excels in all four tracks of the NTIRE19 video restoration and enhancement challenges, surpassing the second place by a significant margin. It also outperforms state-of-the-art methods in video super-resolution and deblurring

TGA [3] introduce an innovative deep neural network that effectively incorporates motion information in a hierarchical fashion, implicitly leveraging complementary details across frames to enhance the reconstruction of missing information in the reference frame. Departing from conventional approaches such as aligning all frames to the reference frame using optical flow or applying 3D convolution to the entire sequence, our proposed method involves dividing the sequence into distinct groups and performing hierarchical information integration. Initially, information integration occurs within each group, followed by integration across groups. Our novel grouping technique results in subsequence groups with varying frame rates, offering diverse complementary information for the reference frame. An attention module is employed to model these distinct types of complementary information, and the groups are intricately fused using both a 3D dense block and a 2D dense block, yielding a high-resolution rendition of the reference frame. In essence, our hierarchical approach enables the handling of diverse motion scenarios and the adaptive utilization of information from groups with different frame rates. For instance, in the case of object occlusion in one frame, the model prioritizes frames where the object is not occluded for more attentive information borrowing.

BasicVSR [1] prioritizes bidirectional propagation for emphasizing long-term and global propagation in restoration. Alignment is achieved through a straightforward flow-based approach at the feature level. Popular choices of feature concatenation and pixelshuffle are used for aggregation and upsampling. Despite its simplicity, BasicVSR excels in both restoration quality and efficiency. Building upon BasicVSR, the authors introduce two innovative components, the **I**nformation-refill mechanism and **cou**pled propagation (**IconVSR**). These additions aim to address error accumulation during propagation and enhance information aggregation.

Redesigning BasicVSR involves the introduction of second-order grid propagation and flow-guided deformable alignment, resulting in the evolution of **BasicVSR++** [5]. This advancement underscores that reinforcing the recurrent framework with augmented propagation and alignment capabilities enables a more effective exploitation of spatiotemporal information across misaligned video frames. The integration of these novel components yields improved performance within comparable computational constraints. Notably, BasicVSR++ surpasses BasicVSR by a substantial 0.82 dB in PSNR while maintaining a similar number of parameters. Furthermore, the versatility of BasicVSR++ extends to various other video restoration tasks.

B. EDVR

1) Paper: EDVR[2] is designed to address various video restoration tasks such as super-resolution and deblurring. The key components of EDVR include alignment module PCD (Pyramid, Cascading, and Deformable convolutions) and fusion module TSA (Temporal and Spatial Attention).

The PCD module utilizes deformable convolutions for feature-level alignment of each adjacent frame with the reference frame and employs a coarse-to-fine alignment strategy to handle intricate and extensive motions. This involves a pyramid structure that initially aligns features in lower scales with coarse estimations, followed by the propagation of offsets and aligned features to higher scales for precise motion compensation. Additionally, PCD cascade an extra deformable convolution after the pyramidal alignment operation to enhance alignment robustness.

The TSA module serves as a fusion mechanism to aggregate information from multiple aligned features. To account for visual informativeness on each frame, temporal attention is introduced by calculating element-wise correlations between the features of the reference frame and each neighboring frame. These correlation coefficients weigh each neighboring feature, indicating its informativeness for reconstructing the reference image. The weighted features from all frames are convolved and fused. Following temporal attention, spatial attention is applied to assign weights to each location in each channel, effectively exploiting cross-channel and spatial information.

The authors participated in all four tracks of the video restoration and enhancement challenges, covering video super-resolution and video deblurring. With the robust alignment and fusion modules, EDVR achieved champion status in all four challenging tracks, showcasing the effectiveness and versatility of our approach. Beyond competition results, EDVR present comparative outcomes on established benchmarks for video super-resolution and deblurring, demonstrating EDVR's superior performance over state-of-the-art methods in these video restoration tasks.

C. TGA

1) Paper: The authors proposed a novel method that can incorporate temporal information in a hierarchical way. Unlike previous frame methods that mostly only use information from adjacent frames, it divides the sequence into several groups with different frame rates. For example, in a group with a dilation of 3, frame n becomes the neighbor of frame $n-3$ and frame $n+3$. This way it establishes a strong relationship between originally non-adjacent frames, making better use of information from non-adjacent frames.

Besides, the authors proposed a new alignment method to futher improve the performance on the data which has large motions. Select a reference frame, use a feature detector to detect points of interest and warp other frames to align with the reference. It is much more efficient and accurate than the traditional optical flow.

2) Experiments: The author adopted Vimeo-90k as training set, sampled regions with spatial resolution 256×256 from high resolution video clips and applied a Gaussian blur with $\sigma = 1.6$ and $4\times$ downsampling (H/4, W/4) to generated low-resolution patches of 64×64. Then they evaluate the proposed method on two benchmarks: Vid4 and Vimeo-90K-T. The model is supervised by pixel-wise L1 loss and optimized with Adam optimizer in which $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Weight decay is set to 5×10^{-4} during training. The learning rate is initially set to 2×10^{-3} and later down-scaled by a factor of 0.1 every 10 epoches until 30 epochs.

We reproduced the experiment using code provided by the authors and obtained similar performance. Additionally, we also evaluate it on the UVG dataset to see if there will be different performance.

3) Result: The result has been shown in TABLE I. In most of the scenes, the TGA has the worst performance.

D. BasicVSR (IconVSR)

1) Paper: For the propagation phase, BasicVSR opts for bidirectional propagation with a focus on long-term and global propagation. In terms of alignment, BasicVSR employs a straightforward flow-based alignment at the feature level. As for aggregation and upsampling, it relies on widely used methods such as feature concatenation and pixel shuffle. Despite its simplicity, BasicVSR attains impressive performance in both restoration quality and efficiency.

Building upon BasicVSR as a foundation, the authors introduce two innovative components – the Information-refill mechanism and coupled propagation (IconVSR). These additions aim to address error accumulation during propagation and enhance information aggregation, contributing to further improvements in the overall performance.

TABLE I
QUANTITATIVE COMPARISON ON BD (PSNR/SSIM)

UVG	BD degradation					
	WavVSR(ours)	**IconVSR+(ours)**	**BasicVSR++**	**IconVSR**	**BasicVSR**	**TGA**
Beauty[15]	35.5378/0.7937	35.5346/0.7938	35.2607/0.7840	35.4982/0.7921	35.5168/0.7929	35.65/0.7981
Bosphorus[2]	42.8018/0.9661	42.7923/0.9660	42.9841/0.9670	42.8700/0.9667	42.9102/0.9666	35.23/0.9015
HoneyBee[8]	39.5415/0.9113	39.4909/0.9109	39.5701/0.9116	39.5393/0.9111	39.5651/0.9115	36.85/0.9370
Jockey[4]	40.4879/0.9297	40.4570/0.9296	40.4598/0.9298	40.4531/0.9297	40.4898/0.9298	39.47/0.9455
ReadySetGo[6]	37.4714/0.9410	37.6073/0.9417	36.8229/0.9341	37.3872/0.9401	37.3592/0.9402	31.45/0.9068
ShakeNDry[9]	39.0294/0.9209	38.9476/0.9203	39.1565/0.9222	39.1208/0.9217	39.0894/0.9214	31.67/0.8264
YachtRide[11]	38.1752/0.9542	38.1086/0.9538	37.9024/0.9534	37.8816/0.9534	38.1740/0.9545	30.74/0.8524
CityAlley[12]	36.9638/0.9170	37.2754/0.9186	37.4685/0.9210	37.5962/0.922	37.0850/0.9188	32.66/0.9022
FlowerFocus[5]	40.3712/0.9161	40.3580/0.9161	40.0936/0.9122	40.3483/0.9157	40.3675/0.9160	42.31/0.9583
FlowerKids[3]	40.3712/0.9584	40.5302/0.9588	40.3712/0.9584	40.4414/0.9586	40.5385/0.9589	32.92/0.9308
FlowerPan[10]	40.5469/0.9589	37.9697/0.9185	37.7833/0.9174	37.9945/0.9192	37.8830/0.9181	31.31/0.8712
Lips[14]	35.7736/0.7929	35.7716/0.7931	35.2058/0.7832	35.9945/0.9192	35.7747/0.7930	38.86/0.8952
RaceNight[13]	36.5994/0.8591	36.5542/0.8585	36.6263/0.8596	36.5825/0.8593	36.6278/0.8593	32.90/0.9074
RiverBank[16]	33.2175/0.8748	32.1073/0.8571	32.5444/0.8673	32.8355/0.8673	32.3042/0.8632	26.44/0.7127
SunBath[1]	48.1034/0.9875	46.2023/0.9866	46.3467/0.9866	46.3426/0.9866	46.3446/0.9866	39.47/0.9804
Twilight[7]	38.2958/0.9285	38.5587/0.9289	37.9982/0.9282	38.3869/0.9287	38.1641/0.928	34.78/0.9295
Average	**38.9555/0.9131**	**38.6416/0.9095**	**38.5371/0.9085**	**38.6880/0.9140**	**38.6371/0.9100**	**34.54/0.8377**

[a] Red and blue colors indicate the best and the seconed-best performance, respectively.
[b] BD degradation means Blur Downsampling with factor 4 (H/4, W/4).

TABLE II
QUANTITATIVE COMPARISON ON BI (PSNR/SSIM)

UVG	BI degradation				
	WavVSR(ours)	**IconVSR+ (ours)**	**BasicVSR++**	**IconVSR**	**BasicVSR**
Beauty[15]	35.6155/0.7967	35.6088/0.7967	35.2877/0.7879	35.4761/0.7912	35.5405/0.7940
Bosphorus[2]	43.6793/0.9683	43.7262/0.9681	43.8598/0.9689	43.7469/0.9688	43.8083/0.9686
HoneyBee[8]	39.7956/0.9131	39.7269/0.9125	39.7473/0.9127	39.5320/0.9092	39.7532/0.9127
Jockey[4]	40.8666/0.9312	40.8541/0.9312	40.7885/0.9306	40.8540/0.9311	40.8884/0.9314
ReadySetGo[6]	39.8096/0.9513	39.5797/0.9501	40.0977/0.9531	39.9648/0.9524	40.0599/0.9523
ShakeNDry[9]	39.5876/0.9247	38.9512/0.9501	39.7310/0.9260	39.6816/0.9255	39.6421/0.9252
YachtRide[11]	39.0836/0.9580	39.5012/0.9240	39.0010/0.9583	38.8438/0.9578	39.1616/0.9585
CityAlley[12]	37.6686/0.9209	38.9512/0.9574	38.3421/0.9257	38.3093/0.9256	37.8486/0.9218
FlowerFocus[5]	40.4919/0.9173	40.4921/0.9173	37.7632/0.8722	40.3612/0.9153	40.3633/0.9152
FlowerKids[3]	42.2422/0.9625	42.0703/0.9620	42.5008/0.9633	42.2751/0.9627	42.3414/0.9628
FlowerPan[10]	39.0845/0.9254	39.0352/0.9246	39.3280/0.9269	39.2454/0.9627	39.2380/0.9264
Lips[14]	35.8373/0.7952	35.8372/0.7954	35.4498/0.7888	35.7714/0.793	35.8291/0.7952
RaceNight[13]	37.1477/0.8619	37.0672/0.8613	37.2928/0.8631	37.1672/0.8624	37.2780/0.8625
RiverBank[16]	32.0810/0.8673	33.0382/0.8686	33.9457/0.8842	33.8444/0.8818	33.6428/0.8792
SunBath[1]	48.0582/0.9874	48.0019/09871	48.1228/0.9875	48.1123/0.9875	48.1034/0.9875
Twilight[7]	38.2386/0.9240	39.7142/0.9324	40.2767/0.9350	39.9555/0.9335	39.9341/0.9335
Average	**39.3305/0.9128**	**39.4367/0.9131**	**39.4710/0.9115**	**39.5713/0.9140**	**39.5895/0.9142**

[a] Red and blue colors indicate the best and the seconed-best performance, respectively.
[b] BI degradation means Bicubic Downsampling with factor 4 (H/4, W/4).

2) Code: Two extensively utilized datasets for training are REDS [6] and Vimeo-90K [7]. In the case of REDS, the authors employ a portion of the REDS4 dataset as their test set, designating REDSval4 as the validation set. The remaining clips are utilized for training. For testing, Vid4 [9], UDM10 [8], and Vimeo-90K-T [7] serve as test sets, in addition to Vimeo-90K. The models are evaluated under 4× downsampling (H/4, W/4) with two degradation methods: Bicubic (BI) and Blur Downsampling (BD). The same downsampling methods are applied to Vimeo-90K. Consequently, three models are trained on REDS4, Vimeo-90K with BI, and Vimeo-90K with BD. Moreover, the evaluation metrics include PSNR and SSIM.

To verify the effectiveness of BasicVSR models, we execute the model trained on the REDS4 [6] dataset on UDM10 (BDx4) [8]. Despite the availability of the pre-trained model, we undertake the preparation of these two datasets and preprocess REDS4. The obtained result is 33.4416/0.9308 (PSNR/SSIM), which closely aligns with their reported result of 33.4478/0.9306.

3) Experiment: The open Ultra Video Group (UVG) [4] dataset consists of 16 versatile 4K (3840×2160) test video sequences captured at 50/120 fps. To begin, all videos with 4K resolution (3840×2160) and 8-bit depth are downloaded. Subsequently, the *ffmpeg* command is employed to convert each video into 600 frames except ShakeNDry and SunBath

(300 frames). Following this, the frames undergo a 4× downsampling process using two degradation methods – Bicubic (BI) and Blur Downsampling (BD). In the final step, the BasicVSR model, trained on REDS4, is applied to the first 32 frames (same as UDM10) from each video, aiming to reduce computational time and resource utilization.

4) Result: Finally, we get results in TABLE I and II. All results are calculated on Y-channel. In contrast to the results reported by BasicVSR, the BasicVSR model trained on REDS4 demonstrates significantly improved performance on the UVG dataset. Our conjecture is that this improvement may be attributed to the higher resolution of the UVG dataset. Noting that the model performs better on the UVG dataset with bicubic degradation (BI) than with blur degradation (BD), it is recommended to prioritize BI in applications. Consequently, an analysis of the performance on UVG(BIx4) is warranted. BasicVSR demonstrates the highest average performance, leading us to organize the performance of each UVG video based on BasicVSR on BI. Upon examination, it becomes apparent that BasicVSR excels most in the SunBath video and performs less optimally in the case of RiverBank.

E. BasicVSR++

1) Paper: BasicVSR++ [5] comprises two impactful revisions aimed at enhancing propagation and alignment. When presented with a video input, initial application of residual blocks facilitates feature extraction from individual frames. Subsequently, a second-order grid propagation scheme is employed to propagate these features, with alignment achieved through the innovative flow-guided deformable alignment method. Following the propagation phase, the aggregated features are utilized in generating the output image through convolution and pixel-shuffling.

2) Result: UVG is evaluated using the BasicVSR++ model trained on REDS, following a similar approach to BasicVSR. The results are shown in TABLE I and II. BasicVSR++ consistently exhibits superior performance, particularly when subjected to BI degradation.

F. IconVSR+ (ours)

1) Analysis: The existing methodologies predominantly rely on the utilization of spatial features (image features) and temporal features (optical flow). To determine whether there exists an positive or negative correlation with model performance in each UVG video, a comparison of spatial and temporal features becomes essential. While the method of comparing spatial features raises questions, it is noteworthy that temporal features can be compared using flow energy. In our analysis, the correlation coefficient obtained for PSNR and flow energy is 0.49, indicating a medium positive correlation showed in Fig. 1. This implies a discernible connection between PSNR and temporal features. Specifically, the lower the flow energy, indicating less motion in the video, the better the model performs in such videos. This result conflicts with our intuition that performance improves when there is minimal motion in the video due to nearly identical frames. Consequently, we interpret this outcome as an indication that the models depend more on temporal features than spatial features. To enhance the model, it is advisable to emphasize the utilization of spatial features in the new model.

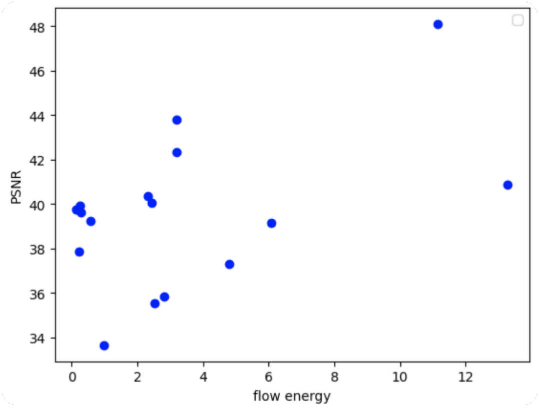

Fig. 1. Temporal feature analysis

2) Model: IconVSR is created by replacing the EDVR-M feature extractor of information-refill part in IconVSR with TGA feature extractor. There are two reasons of such design. First, average performance of IconVSR with EDVR-M on UVG is the best with BD degradation and the second best with BI degradation. EDVR-M is a mini version of EDVR. Second, both TGA and EDVR more depend on spatial feature and without optical flow, and TGA performance is better than EDVR, according to the TGA paper.

As shown in Fig. 2, a TGA feature extractor is used to extract deep features from a subset of input frames (keyframes) and their respective neighbors. The extracted features are then fused with the aligned features.

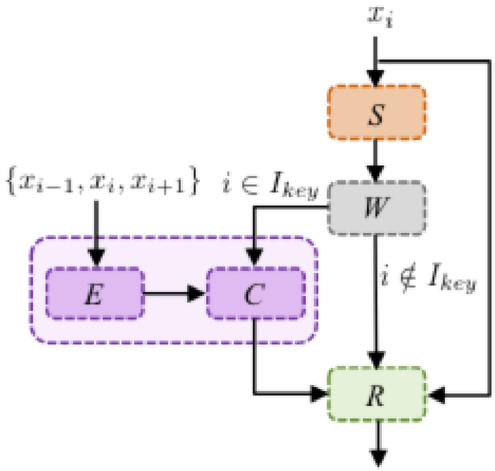

Fig. 2. Information-Refill. E and C correspond to the feature extractor and convolution, respectively. S, W , and R refer to the flow estimation module, spatial warping module, and residual blocks, respectively.

3) Experiment: We utilize the REDS [6] datasets for training, with REDSval4 serving as the validation set. The

TABLE III
QUANTITATIVE COMPARISON

	WavVSR	IconVSR+	BasicVSR++	IconVSR	BasicVSR
Params (M)	6.291	-	7.323	8.695	6.291
FLOPs (T)	2.089	-	2.217	2.865	2.089
Runtime (ms)	135	165	138	138	118

[a] Blanked entries correspond to results unable to be reported.
[b] The FLOPs is computed on a shape (5, 3, 256, 256).
[c] The runtime is computed on an LR size of 960 × 540.
[d] WavVSR has the same architecture as BasicVSR.

remaining clips are employed for the training process. UVG [4] is employed as the test set, and our models undergo testing with 4× downsampling (H/4, W/4) using two degradation methods – Bicubic (BI) and Blur Downsampling (BD).

We employ a pre-trained SPyNet as our flow estimation module. The optimization process utilizes the Adam optimizer and follows a Cosine Annealing scheme. The initial learning rates for the feature extractor and flow estimator are established at 1×10^{-4} and 2.5×10^{-5}, respectively. The learning rate for all other modules is set to 2×10^{-4}. The total number of iterations is set to 30K, with the weights of the feature extractor and flow estimator remaining fixed for the first 500 iterations. The batch size is set to 8, and the patch size of input LR frames is 64×64. Charbonnier loss is employed due to its superior handling of outliers, leading to improved performance over the conventional L2 loss.

4) Result: The outcomes are displayed in TABLE II and II. IconVSR+ achieves the highest performance in YatchtRide and CityAlley, securing the second-best performance in Beaty and Lips. Notably, BasicVSR performs poorly on these four UVG videos. This implies that IconVSR has indeed enhanced performance, especially on UVG videos where the referenced models struggle. Despite having only 30K training iterations due to time and GPU constraints, compared to the 300K iterations for BasicVSR and BasicVSR++, IconVSR exhibits commendable performance.

G. WavVSR (ours)

1) Analysis: Most of the existing video super-resolution methods are based on the loss function of rgb domain to supervise the neural network, i.e. minimizing the mean-square error(MSE) of the rgb value of the pixels. This loss function is intuitive since the evaluation criteria of the model are also based on rgb domain. However, as argued in many papers [10], merely minimizing MSE loss can hardly capture high-frequency texture details and often produces over-smooth results. Like the forest in the RiverBank scene in the UVG dataset, the BasicVSR produces blurry results.As texture details can be depicted by high-frequency wavelet coefficients, we transform super resolution problem from original image space to wavelet domain and introduce wavelet-based losses to help texture reconstruction. Inspired by PatchGAN [11], patch-wise supervision can pay better attention to local details than image-wise supervision. Therefore, we believe that introducing patch-wise wavelet-based loss can improve the BasicVSR performance on the scene with a large amount of high frequency details.

2) Model: Before applying the wavelet-based loss to compare the prediction and the ground truth, we first divide these two images to 4×4 patches with equal size. The wavelet-based loss consists of two parts: wavelet prediction loss and texture loss. The former one is the MSE in wavelet domain, defined as

$$l_{wavelet}(\hat{C}, C) = ||(\hat{C} - C)||^2$$

where C and $\hat{C}$ denoted the ground-truth and inferred wavelet coefficients respectively. We adopt a similar idea with FSSR [12], give up the LL part and only imposes loss in the high-frequency space. The texture loss is designed to prevent high-frequency wavelet coefficients from converging to zero, defined as

$$l_{texture}(\hat{C}, C) = max(||\hat{c}||^2 - ||c||^2)$$

where $c \in C$ and $\hat{c} \in \hat{C}$ denoted the pixels in the spectral graph. The texture loss focus on the pixel with the worst prediction and hence prevents the degradation of texture details. The unified loss function is defined as follows

$$l_{total} = \alpha l_{mse} + \beta l_{wavelet} + \gamma l_{texture}$$

where the l_{mse} is the original loss used in the BasicVSR, α,β and γ are the balance parameters.

3) Experiment: Be consistent with the original paper, We utilize the REDS [6] datasets for training, with REDSval4 serving as the validation set and the remaining clips are employed for the training process. And be consistent with the experiment of IconVSR+, UVG [4] is employed as the test set, and our models undergo testing with 4× downsampling (H/4, W/4) using two degradation methods – Bicubic (BI) and Blur Downsampling (BD). We employ a pre-trained SPyNet as our flow estimation module. The optimization process utilizes the Adam optimizer and follows a Cosine Annealing scheme. The initial learning rates for the feature extractor and flow estimator are established at 1×10^{-4} and 2.5×10^{-5}, respectively. The learning rate for all other modules is set to 210^{-4}. The total number of iterations is set to 30K, with the weights of the feature extractor and flow estimator remaining fixed for the first 500 iterations. The batch size is set to 8, and the patch size of input LR frames is 64 × 64. The balance parameter α is set to 0.5, β is set to 0.25, γ is set to 0.25.

4) Result: The outcomes are displayed in TABLE II and II. Evaluated on the 16 scenes in the UVG data with BD degradation, WavVSR achieves 4 best performances and 4 second-best performances among the 6 models, and it outperforms the BasicVSR in 8 scenes. When with BI degradation, WavVSR achieves 3 best performances and 2 second-best performances and it outperforms the BasicVSR in 4 scenes. It is worth mentioning that, WavVSR achieves 31.1258db in the validation dateset, which is the best performances among the existing models: BasicVSR(30.17), IconVSR(w/o refill)(30.38) and IconVSR(30.45).

III. Summary

Many current approaches consist of four interconnected components: propagation, alignment, aggregation, and upsampling, predominantly leveraging spatial and temporal features. In our modification of IconVSR, we enhance the aggregation aspect by substituting the EDVR-M extractor with the TGA extractor, aiming to maximize the utilization of spatial features. Besides, we introduce the patch-wise wavelet-based loss to supervise the model in the frequency domain, aiming to capture the high frequency details. These refinement results in the creation of superior models, IconVSR+ and WavVSR.

In our future endeavors, we intend to further train IconVSR+ for additional iterations. Additionally, we aspire to devise a method for comparing videos, particularly focusing on spatial features, and analyze their impact on the performance of Video Super-Resolution (VSR).

References

[1] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond." arXiv, Apr. 07, 2021. doi: 10.48550/arXiv.2012.02181.

[2] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video Restoration with Enhanced Deformable Convolutional Networks." arXiv, May 07, 2019. doi: 10.48550/arXiv.1905.02716.

[3] T. Isobe et al., "Video Super-resolution with Temporal Group Attention." arXiv, Jul. 21, 2020. doi: 10.48550/arXiv.2007.10595.

[4] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," in Proc. ACM Multimedia Syst. Conf., Istanbul, Turkey, June 2020.

[5] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5972–5981.

[6] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and superresolution: Dataset and study. In CVPRW, 2019.

[7] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. IJCV, 2019.

[8] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In ICCV, 2019.

[9] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. TPAMI, 2014.

[10] Huaibo Huang, Ran He, Zhenan Sun and Tieniu Tan. Wavelet-SRNet: A Wavelet-based CNN for Multi-scale Face Super Resolution.ICCV1 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. https://arxiv.org/abs/1611.07004.

[12] Manuel Fritsche, Shuhang Gu, Radu Timofte. Frequency Separation for Real-World Super-Resolution. https://arxiv.org/abs/1911.07850.