A Study of Video Super-Resolution Method Using Video Coded Data as Training Data

Remina Yano
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
yano.remina@toki.waseda.jp

Yun Liu
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
yuinn@asagi.waseda.jp

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

Takuya Suzuki
SHARP Corporation
Chiba, Japan
takuya.suzuki@sharp.co.jp

Takeshi Chujoh
SHARP Corporation
Chiba, Japan
chujoh.takeshi@sharp.co.jp

Tomohiro Ikai
SHARP Corporation
Chiba, Japan
ikai.tomohiro@sharp.co.jp

***Abstract*—This paper presents a study of video super resolution method by applying Deformable Convolution network to coded video. It is reported the effectiveness of using coded video as training data, and numerical and visual results of coded fine-tuned model. From those results, it is discussed the relationship between about characteristic of training data and especially in video's framerate. There it is shown that video sequence which has similar framerate with training data can perform higher PSNR, since low framerate means the motion between frames is large and Deformable Convolution can learn that large motion.**

Keywords—video super resolution, Deformable Convolution, Versatile Video Coding (VVC)

I. Introduction

In 2021, 4K/8K UHDTV has already been in operation, and devices for outputting these high-resolution images as well as efficient coding methods are being explored. On the other hand, super-resolution technology has been widely studied as a means to enhance the resolution and quality of images and videos.

In this paper, we describe the targeted super-resolution method and its effectiveness by discussing following three topics. First, we use the video super-resolution network named Temporally-Deformable Alignment Network for Video Super-Resolution (TDAN) [1] to verify super-resolution for video coded by the next-generation video coding standard, Versatile Video Coding (VVC). Second, we show the effectiveness of fine-tuning applied to TDAN using VVC-encoded / degraded video with reduced resolution instead of the clean bicubic processed one for training data. Finally, we focus on the frame rate of the training and test data to make observations about the super-resolution results.

II. Related Works

A. Deformble Convolution

Deformable Convolution[2] is a technique for optimizing the convolutional position and distance. Since it learns the offset in the convolution neural network as well as weight and bias, the receptive field becomes flexible depending on the convolution point of the object, as shown in Fig. 1. In the field of video super-resolution, Deformable Convolution can perform temporal alignment more smoothly, i.e., motion estimation and motion compensation in a single step. Deformable Convolution has been getting attention as an effective method for reducing occlusions and artifacts.

Fig. 1. Example of Deformable Convolution. Yellow dots show Convolution points and red dots show their receptive fields. The left image shows the case where the convolution point is in the background (sky), and the right image shows the case where the convolution point is in an object (car).

B. TDAN

TDAN, shown in Fig. 2, is a video super-resolution network using Deformable Convolution. TDAN estimates the offsets, which could be considered as the motion displacement between frames, using the two frames before and after target low-resolution frame, and they create the target high-resolution frame using the estimated offsets. Although the offsets learning of Deformable Convolution and the super-resolution module learning use different loss functions, TDAN actually uses the sum of each loss function, which enables end-to-end learning.

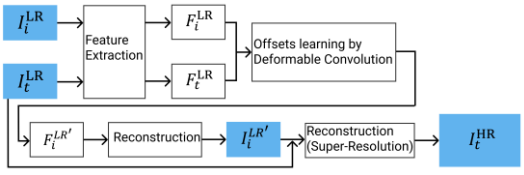

Fig. 2. Block diaglam of TDAN proceccsing. Subscript t indicates the time of target frame and i indicates the time of sorrounding frame. In our experiments, we set i to t-2, t-1, t+1, and t+2.

III. Proposed Method

In this research, we apply TDAN super-resolution network (scale = 2) to video data coded and reduced by VVC, and

verify output video quality by using image evaluation index PSNR.

In our training process, we first create anchor model named TDANx2 using Vimeo Super- Resolution dataset et al. [3]. Then, we fine-tune the TDANx2 using SJTU [4] and UVG [5] dataset to create super-resolution models named A-dec1 and A-dec2. In fine-tuning process, we used 1/2 reduced resolution VVC coded video by VVC Test model (VTM) utilizing the regerence picture resampling (PRR) encoder option. A summary of the data used to train models A-dec1 and A-dec2 is shown in Table I.

TABLE I. CODED DATA 1 AND 2 USED TO CREATE MODELS A-DEC1 AND A-DEC2

frame rate (fps)	**Number of Sequences**	
	Coded data 1 (A-dec1)	***Coded data 2 (A-dec2)***
30	3	14
50	8	8
120	7	7

IV. RESULTS AND CONSIDERATION

We evaluate our model by using JVET Common Test Conditions (CTC) for SDR. [6]. The experimental result of this study is shown in Table Ⅱ. In Table Ⅱ, frame rate is shown in parentheses after sequence name.

It can be seen that model A-dec1 and A-dec2 get high PSNR score while TDANx2 slightly over bicubic interpolation. In the top three sequences, the proposed method A-dec2 gives the best results, but for the BQSquare sequence, A-dec1 gives the best results. The original resolution of all four sequences is 416x240 pixels, and the input's resolution converted to 208x120 pixels by VTM. Fig. 3 and Fig. 4 shows images of the experimental results for the RaceHorsesD and BQSquare sequences.

TABLE II. PSNR(DB) FOR EXPERIMENTAL RESULTS OF EACH MODEL

Sequence	**Model**			
	Bicubic	***TDANx2***	***A-dec1(ours)***	***A-dec2(ours)***
RaceHorsesD (30fps)	23.839	23.910 (+0.071)	24.077 (+0.238)	24.584 (+0.745)
BasketballPass (50fps)	25.552	25.369 (-0.183)	25.633 (+0.081)	25.839 (+0.287)
BlowingBubbles (50fps)	22.778	22.800 (+0.022)	23.008 (+0.230)	23.105 (+0.327)
BQSquare (60fps)	20.602	20.603 (+0.001)	21.123 (+0.521)	20.578 (-0.024)

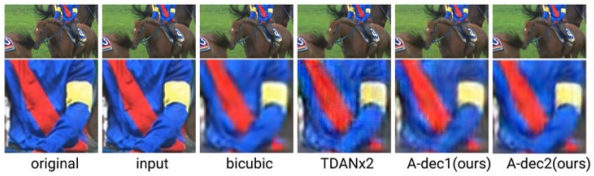

Fig. 3. From left to right, above images show correct image, input image, bicubic interpolation, TDANx2, A-dec1 and A-dec2 results of the sequence RaceHorsesD, respectively.

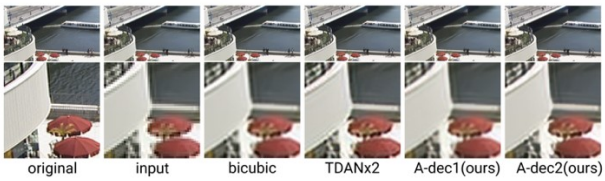

Fig. 4. From left to right, above images show correct image, input image, bicubic interpolation, TDANx2, A-dec1 and A-dec2 results of the sequence BQSquare, respectively.

Since the BQSquare sequence is a landscape scenery while other sequences have some actions, it has a feature that the frame rate is higher than that of other sequences and there is little motion between frames. From Table I it can be seen that A-dec2 contains 14 sequences with a frame rate of 30 fps. It means that A-dec2 is created from data with large motion between frames. Therefore, it is considered that the PSNR value of A-dec2 has become smaller in the BQSquare sequence, which has low motion between frames. Therefore, we can conclude that, if the video has a framerate similar to the training data, we will get a high-resolution video with good quality since deformable convolution can learn those motions.

V. CONCLUSION

We propose a method to improve the visual quality of video super-resolution of coded video to have less degration by applying a super-resolution network (TDAN) with appropriately selected datasets. From the experimental results, it is confirmed that the video with the similar motion between frames as the training data leads to high quality.

A future task is to verify the effectiveness of the proposed method with test data under the same conditions as the video used in the new 4K/8K broadcasting. In addition, it is considered necessary to devise super-resolution network more flexible for various motion videos by applying a model learned according to the framerate of the target image and the amount of movement between frames.

REFERENCES

[1] Y.Tian, Y.Zhang, Y.Fu and C.Xu, "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2020), pp.3357-3366, June 2020

[2] J.Dai, Y.Xiong, Y.Li, G.Zhang, H.Hu, and Y.Wei, "Deformable Convolutional Networks," 2017 IEEE International Conference on Computer Vision (ICCV2017), pp.764-773, Oct. 2017

[3] T.Xue, B.Chen, J.Wu, D.Wei, and W.T.Freeman, "Video enhancement with task-oriented flow," International Journal of Computer Vision, vol. 127, no. 1, pp.1106-1125, Feb.2019.

[4] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K Video Sequence Dataset," the Fifth International Workshop on Quality of Multimedia Experience (QoMEX2013), pp.34-35, July 2013.

[5] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," The 11th ACM Multimedia Systems Conference (MMSys20), pp.297-302, June 2020.

[6] F.Bossen, J.Boyce, K.Sühring, X.Li, and V.Seregin, "JVET common test conditions and software reference configurations for SDR video," Joint Video Experts Team Document, JVET-N1010, May 2019.