Video Super-Resolution based on selected Alignment Methods with New Evaluations

209 AS Final Report

Zichao Xian (205626091)
Yuanlong Chen (805846962)
Boya Ouyang (005037574)

1. Introduction

Compared to single-image super-resolution, which focuses on the intrinsic properties of a single image for the upscaling task, video super-resolution (VSR) poses an extra challenge as it involves aggregating information from multiple highly-related but misaligned frames in video sequences.. One prevalent approach is the sliding-window framework [1], where each frame in the video is restored using the frames within a short temporal window. In contrast to the sliding window framework, a recurrent framework attempts to exploit the long-term dependencies by propagating the latent features. In general, these methods [2] allow a more compact model compared to those in the sliding window framework. Nevertheless, the problems of transmitting long-term information and aligning features across frames in a recurrent model remain formidable.

In this project we adopt a bidirectional propagation scheme to maximize information gathering, and an optical flow-based method to estimate the correspondence between two neighboring frames for feature alignment. By simply streamlining these propagation and alignment components with the commonly adopted designs for aggregation and upsampling , our approach outperforms the baseline model without feature alignment.

2. Related work:

2.1 Recurrent Networks.

The recurrent framework is a popular structure adopted in various video processing tasks such as super-resolution], deblurring,, and frame interpolation.. For instance, RSDN [3] adopts unidirectional propagation with a recurrent detail structural block and a hidden state adaptation module to enhance the robustness to appearance change and error accumulation. Chan et al. [4] propose BasicVSR. The work demonstrates the importance of bidirectional propagation over unidirectional propagation to better exploit features temporally. In addition, the study also shows the advantage of feature alignment in aligning highly relevant but misaligned features

2.2 Alignment method

For the traditional Video Super Rsesolution (VSR) approaches, two frameworks are often used to process the video. One is called sliding-window, another is recurrent. The sliding-window framework predicts the optical flow between low-resolution frames and performs the spatial warping for the alignment. With the development of it, approaches restoring to more sophisticated and implicit alignment, such as TDAN uses deformable convolutions (DCNs) to align different frames at the feature level, EDVR adopts DCNs in a multi-scale fashion for more accurate alignment, and DUF makes use of dynamic upsampling filters for handling motions implicitly. Meanwhile, the recurrent framework such as RSDN uses detail-structural block and adapt a hidden state module to enhance the robustness to accumulate error and change appearance, and RRN uses a residual mapping between layers by identity skipping connections for ensuring a fluent information flow and protect the texture information for long time. [5]

Recently, the research showed that with spatial alignment in not only the image level, but also in the feature level, it would generate a marked improvement. In the BasicVSR

project, the researchers make comparisons between image warping and feature warping. Due to the inaccurate optical flow estimation, the warped images suffer from blurriness and incorrectness. 0.17db in PSNR is observed in the experiments, which confirms the importance of using the alignment method to the feature level instead of only the image level. [6]]

2.3 Deformable Alignment:
Several works [7] employ deformable alignment. TDAN performs alignment at the feature level using deformable convolution. EDVR further proposes a Pyramid Cascading Deformable (PCD) alignment with a multi-scale design. Recently, Chan et al. [8] analyze deformable alignment and show that the performance gain over flow-based alignment comes from the offset diversity.

Motivated by [8], we adopt deformable alignment but with a reformulation to overcome the training instability.. Our flow-guided deformable alignment is different from offset-fidelity loss.. The latter uses optical flow as a loss function during training. In contrast, we directly incorporate optical flow into our module as base offsets, allowing a more explicit guidance, both during training and inference.

2.4 Evaluation of recovered images
With more and more models such as BasicVSR, BasicVSR++ and RealBasicVSR are designed to do super resolution with using many different datasets such as REDS4 and Vimeo90K, the evaluations of the recovered results are mostly using Peak-Signal-to-Noise Ratio (PSNR) and Structural Similarity Indexing Method (SSIM) as the parameters of quality evaluation. Since PSNR calculates the power effect of distorting noise to the maximum possible signal [9], it deviates from people's view sometimes because it focuses more on the pixel difference of the picture itself, which is too objective to make people recognize the effect sensitively. Meanwhile, SSIM is usually used to evaluate the results between the original and the recovered images, as it focuses more on the structural information such as luminance and contrast of the images [9]]. However, since people are more sensitive to the luminance and contrast of a restored image, it is too subjective to simply use SSIM as the evaluation standard.

Another evaluation parameter between ground truth and the recovered image is called Perceptual Index (PI), which first came out in the 2018 Perceptual Image Restoration and Manipulation (PIRM) challenge. It is a no-reference quality metric for natural and single image super resolution, which is based on three hypothesized statistical properties (local frequency variations, global frequency variations, and spatial discontinuity) to quantify and evaluate the quality of SR images [10]. However, PI value is given by scoring and training without ground truth, so it is subjective as well.

Additionally, since people are more sensitive to the color and brightness of an image, there is a parameter called Colorfulness from a GitHub project named percuss [11]. It is a method from node-opencv. It comprehensively concludes the color variety, structure complexity, and brightness information to give a direct score to evaluate how colorful the image is. It is based on the centers of BGR-centers distance cost matrix in LUV distance space, which calculates the sum of distance of different cubes in the matrix. In this case, the calculated colorfulness of the image is helpful for people to be sensitive to evaluate the quality of an image, and it is objective because of totally machine calculated results without adding neuro networking training.

The major limit of the evaluation parameters mentioned above is that these are not comprehensive enough if considered separately, either too objective or too subjective, or deviate from people's sensitivity. In this case, a more overall evaluation method should be considered, which should be a score that concludes all the subjective, objective and people sensitive factors.

3. Technical Approach

What we have is just a 1080Ti gpu with 12G memory in the lab. Due limited computational resources, we set the num of frames as 2. For the flow-based DCN alignment method, the least number of frames is 2. While the optical flow method only needs 2 frames to generate flows.

We designed three tests. The first is a Basicvsr network with optical-flow based alignment. The second is a Basicvsr network with flow-guide DCN based alignment. The last is also a Basicvsr whose alignment function is replaced by a very simple linear transformation layer, which means no alignment.We want to put those methods on the top of Basicvsr model[12].

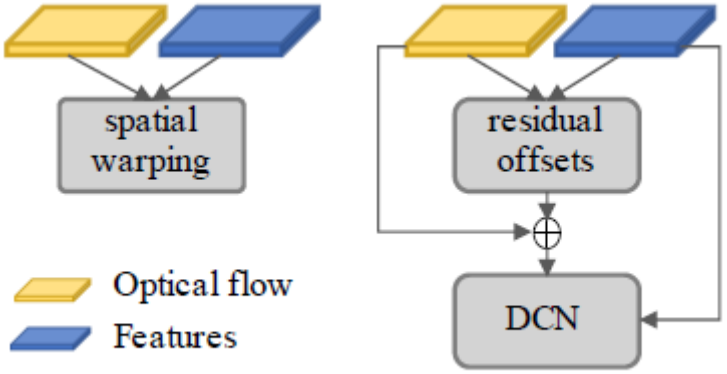

Figure 1: optical flow alignment(left) and flow-guide DCN based alignment (right)

4. Evaluation and Results

4.1 Results of alignment methods analysis

Table 1 shows the training time for DCN, optical flow and linear transformation. It takes the longest time (33 hours) to train DCN due to its complex structure. While it only takes 6 hours to finish training without feature alignment.

The SSIM and PSNR for DCN, optical flow and linear transformation is shown in figure 2 and figure3. It shows that optical flow is the most optimal alignment method with PSNR increasing with training time. In contrast, the PSNR doesn't show improvement with training time for linear and DCN alignment.

Alignment method	DCN	optical flow	linear trans
Train time	about 33 hours	about 8 hours	about 6 hours

Table1. Training time for different alignment methods..

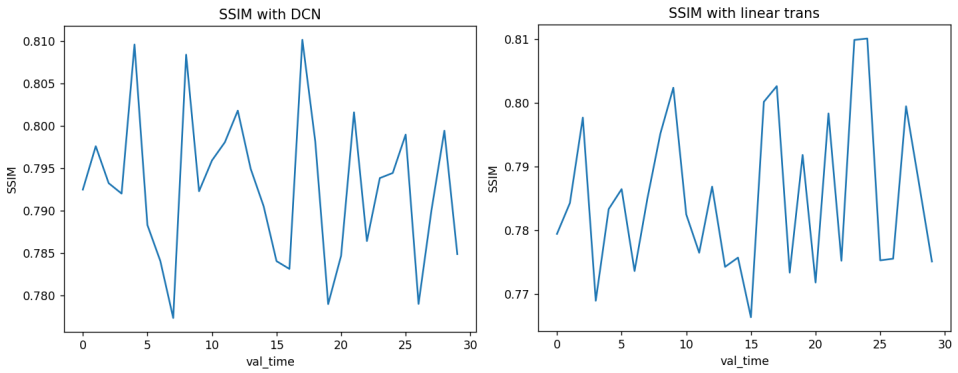

Figure 2. SSIM as function of training time for DCN and linear transform.

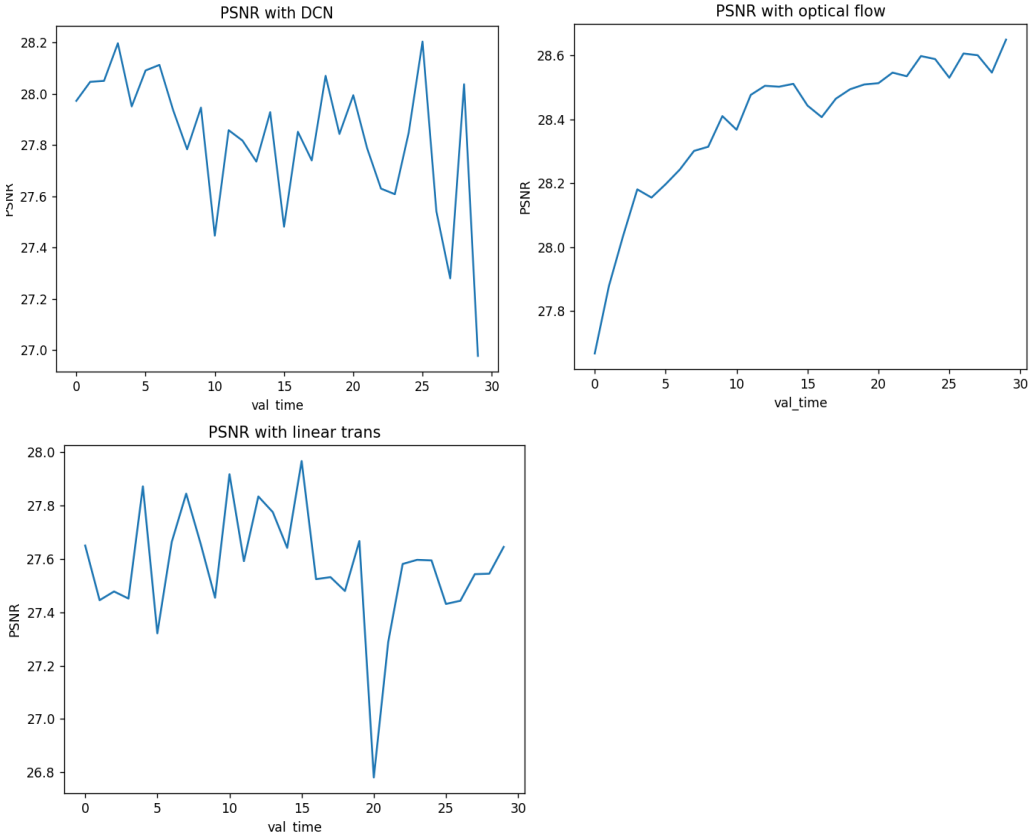

Figure 3. PSNR as a function of training time for DCN, optical flow and linear transform.

4.2 New evaluation: Overall Score

As mentioned above, mostly the results of the super resolution process are evaluated by checking PSNR and SSIM values. However, due to the properties of PSNR and SSIM, PSNR is subjective since it is a result completely calculated by machine according to the pixel difference of each image, and SSIM is objective since it focuses more on the structure and luminance information, which is more apparent to people's views. Both are higher to get better results.

$$PSNR = 10log_{10}(\frac{(L-1)^2}{MSE}) = 20log_{10}(\frac{L-1}{RMSE})$$

[5]

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

[14]]
Meanwhile, PI is another value to evaluate the results, which was firstly given from the 2018 PIRM challenge. Since PI is calculated by analyzing each image itself without any ground truth, it is objective as it is calculated from hypothesized quantifying properties' quality. It is smaller to get better results.

$$\mathrm{PI} = \tfrac{1}{2}\left((10 - \mathrm{Ma}) + \mathrm{NIQE}\right)$$

[9]]

Besides, another value called Colorfulness is also used to evaluate how colorful an image is. By considering the colors and components of an image, colorfulness is subjective. It is higher to be more colorful.

According to all the evaluation values mentioned above, the idea of giving a simple and clear overall score that concludes both the objective and subjective properties come up. Due to the relations between parameter and state of results, it is listed out to check in a table as follows:

PSNR	Subjective	↑ - Better
Colorfulness	Subjective + People sensitive	↑ - More colorful
PI	Objective	↓ - Better
SSIM		↑ - Better

According to the effects of increase and decrease to the evaluation of results, it is considered to find an algorithm to consider the objective and subjective factors simultaneously to calculate a value to be the overall score to the SR results evaluation. And since people are more sensitive to the color and brightness information of an image, the mentioned parameter Colorfulness should be in the calculation of the overall score. In this case, it comes up an equation to conclude all the observation and considerations above, which is shown as follows:

Overall score = PSNR * SSIM * Colorfulness / PI
(↑ - Better)

5. Discussion and Conclusions

In conclusion, we adopt a bidirectional recurrent neural network with feature alignment to perform video super resolution tasks. We compare three alignment methods:linear transformation, optical flow and flow-guided deformable alignment. After comparing PSNR as a function of training time for those alignment methods, we conclude that optical flow alignment is the most promising one showing improvement with training time. Future direction can be using deformable ConvLSTM structure for One-Stage Space-Time Video Super-Resolution.ConvLSTM structure can aggregate temporal and spatial information in a single step by exploring intra-relatedness between temporal interpolation and spatial super-resolution. It can be more effective than existing two-stage

networks.

6. References

[1] Jose Caballero, Christian Ledig, Aitken Andrew, Acosta Alejandro, Johannes Totz, Zehan Wang, and Wenzhe Shi. Realtime video super-resolution with spatio-temporal networks and motion compensation. In CVPR, 2017. 5

[2] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In CVPR, 2021. 1, 2, 3, 5, 8, 10

[3] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In AAAI, 2021. 2, 4, 8

[4] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In ICIP, 1994. 5, 10

[5] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In CVPR, 2018. 2

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, 2017. 2, 4

[7] Damien Fourure, Remi Emonet, ´ Elisa Fromont, Damien ´ Muselet, Alain Tremeau, and Christian Wolf. Residual conv- ´ deconv grid network for semantic segmentation. In BMVC, 2017. 2

[8] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In ICCVW, 2019. 1, 2, 5

[9] Sara, U., Akter, M. and Uddin, M.S. (2019) Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. Journal of Computer and Communications, 7, 8-18. Available: https://doi.org/10.4236/jcc.2019.73002

[10] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, Lihi Zelnik-Manor; Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0-0. Available: https://openaccess.thecvf.com/content_ECCVW_2018/papers/11133/Blau_2018_PIRM_Challenge_on_Perceptual_Image_Super-resolution_ECCVW_2018_paper.pdf

[11] Datta, R., Joshi, D., Li, J., Wang, J.Z. (2006). Studying Aesthetics in Photographic Images Using a Computational Approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3953. Springer, Berlin, Heidelberg. Available: https://doi.org/10.1007/11744078_23. Github Project: https://github.com/piercus/colorfulness

[12] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, Chen Change Loy, "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond," , 18 February 2012. [Online]. Available: https://arxiv.org/pdf/2012.02181.pdf.

[13] imavijit, "Python | Peak Signal-to-Noise Ratio (PSNR)," GeeksforGeeks, 06 February 2020. [Online]. Available: https://www.geeksforgeeks.org/python-peak-signal-to-noise-ratio-psnr/.

[14] WIKIPEDIA, "Structural similarity," , 8 May 2022. [Online]. Available: https://en.wikipedia.org/wiki/Structural_similarity.