BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Neural Networks for Video Quality Enhancement

Matej Sirovatka*

Abstract

Low quality video is a common problem that can be an obstacle in many areas, such as security, medical imaging, self driving cars, games, etc. There are many different reasons for visual imperfections, such as camera movement or bad light conditions. To compensate for those visual imperfections, different methods are used, such as using neural networks. Task of doing so is called video super-resolution, it is a process of increasing the spatial dimensions of the video frames, while removing the visual imperfections and noise. This paper proposes a novel architecture of a neural network, that is employed for video super-resolution. A common problem in current neural networks for video super-resolution is how to align features from the neighbouring frames. The proposed solution is based on using precalculated optical flow in combination with deformable convolution layers to align these features, which can then be further processed by next layers. The solution incorporates this module inside an adapted U-Net [1] like architecture, which shows an improvement of around 0.9 dB (total 28.28 dB) in peak signal-to-noise ration and of around 0.03 (total 0.82) in structural similarity index compared to a basic U-Net architecture on a task of 4x video super-resolution on a validation dataset.

*xsirov00@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

There are multiple areas where high quality video is important, for example in self driving cars, having high quality video is crucial for further object detection and classification, also in medical imaging and security footage. However this task proved to be quite challenging, due to the wide variety of visual imperfections that can appear in a video.

A common problem in a video super-resolution is motion blur, to address this problem, current deep learning based solutions align features from multiple neighbouring frames to obtain high quality details of them. Different solutions have been proposed to solve this issue, such as attention in Visual Restoration Transformer (VRT) by Liang *et al.* [2], but a common issue of those is their computational complexity.

This paper proposes a novel architecture to align neighbouring features using deformable convolutions [3], as proposed by Dai *et al.*, and optical flow estimation. The aim is to use the optical flow between reference and neighbouring frames, to align the neighbouring features towards the reference ones. The primary advantage of this method is relatively low computational complexity, effectively adding only N deformable convolutions for $N+1$ input frames, for each level of U-Net architecture.

2. Proposed solution

The proposed solution is using an U-Net like architecture, to process multiple neighbouring frames of the input video, to increase their spatial resolution 4 times, and remove visual imperfections. During the feature extractions, an encoder is used to extract features of decreasing spatial, but increasing channel dimensions. These features are then used as an input to an alignment module based on deformable convolutions together with precalculated optical flow, which outputs aligned features that are then further processed and upscaled to produce the resulting high quality frame. The whole pipeline is as shown in **Figure 1**. To obtain optical flow, Recurrent All-Pairs Field Transforms for Optical Flow (RAFT) [4] network, as proposed by Teed *et al.* is used. Feature alignment, as shown in **Figure 3**, is incorporated on all of the levels of U-Net architecture, to align different types of features. These features are then concatenated and fused together using a simple convolution layer,

dividing their channel dimensionality by 3, effectively creating a single feature map, that can be then used as an input to each level of the U-Net decoder.

Optical flow, which is a dense displacement field (f_1, f_2) mapping each pixel (u, v) in frame F_i, to its corresponding location $(u', v') = (u + f_1(u), v + f_2(v))$ in frame F_j is used as an offset input to deformable convolution. This is done for each of the neighbouring frames, with optical flow from the reference one towards the neighbouring. With this approach, receptive field of the convolution kernel is shifted towards the respective feature in the neighbouring frame, creating higher quality feature representation as can be seen in **Figure 2**.

3. Training and Evaluation

Evaluating the quality of a video is a tricky task, mainly because of the subjectivity. Different people may view the resulting videos differently, despite this, there are few metrics that help evaluating the video quality. These metrics are primarily used for images, but can be adapted to video by averaging the results over each frame. There were 3 main metrics used in training and evaluation, those being:

1. **Mean Squared Error (MSE)** - this metric is used only as a loss function for the neural network to learn, it effectively measures the mean squared distance between values of each pixel in between 2 images.
2. **Peak Signal-to-noise Ratio (PSNR)** - this metric quantifies the amount of noise introduced during image processing between 2 images. The measurement unit are decibels, providing a logarithmic scale for the comparison of the original and processed image. Higher values of PSNR indicate better results.
3. **Structural Similarity Index (SSIM)** - another metric used for image quality evaluation, calculated using luminance, contrast and structure. The main idea is that the pixels in the images are not independent, but connected. Each pixel is closely related to its neighbouring pixels. This metric considers image degradation as change in structural information, and the resulting value is an index of similarity of 2 images.

4. Experiments

The main purpose of the experiments was to prove, that the proposed solution improves the quality of the resulting image. This was done by comparing it to a single frame U-Net architecture adapted for video super-resolution and bilinear interpolation upscaling. The resulting metrics are as shown in **Table 1** and example output images can be seen in **Figure 4** and **Figure 5**. Further work was done to improve the architecture of the feature alignment module, such as using the fused features from the lower levels to refine the features at the current level further. This surpassed the self-standing levels slightly, increasing the metrics. Another experiments were done, using different U-Net encoder architectures, or decreasing the number of U-Net levels.

5. Future work

Future experiments with the feature alignment module could prove to yield better results. A big place for improvement is in the feature fusion part, where currently only single convolution layer is used. Using attention could yield better results, but with increasing computational complexity. Also incorporating this module into diffusion models could show provide an interesting results, while using the previously output high quality frame to keep the output video consistent. Further research into different U-Net encoder architectures, and number of levels could be introduced. Another place for an improvement is in the upscaling part of the neural network, where using different layers than current pixel shuffle layers might prove beneficial.

Acknowledgements

I would like to thank my supervisor Ing. Michal Hradiš, Ph.D. for their help and advice.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[2] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer, 2022.

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks, 2017.

[4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.