Temporal Consistently Edge-Informed Video Super-Resolution (Edge-VSR)

Ayush Singh
Indian Institute of Technology
Dhanbad, Jharkhand, India
ayush.s.18e0204@cse.iitism.ac.in

Mehrzan Ebrahimi
Imaging Lab
Faculty of Science, Ontario Tech University
Oshawa, Ontario, Canada
mehrzan.ebrahimi@ontariotechu.ca

Abstract

Resolution enhancement of a given video sequence is known as video super-resolution. We propose an end-to-end trainable video super-resolution method as an extension of the recently developed edge-informed single image super-resolution algorithm. A two-stage adversarial-based convolutional neural network that incorporates temporal information along with the current frame’s structural information will be used. The edge information in each frame along with optical flow technique for motion estimation among frames will be applied. Promising results on validation datasets will be presented. All of the video results in this paper are accessible at: https://bit.ly/2WHa1HT

1 Introduction

Naturally, there is always a demand for higher quality and higher resolution images or videos. The level of image detail is crucial for the performance of many computer vision algorithms. When resolution cannot be improved either because of cost or hardware physical limits, one can resort to resolution enhancement algorithms. Even when superior equipment is available, such algorithms provide an inexpensive alternative. The process of producing a high-resolution (HR) image given a single low-resolution (LR) image is called Single Image Super-Resolution (SISR) [1][7]. The problem of recovering a HR video from a given LR one is known as Video Super-Resolution (VSR) [8][14].

Resizing of an image or video does not translate into an increase in its resolution. In fact, resizing should be accompanied by approximations to frequencies higher than those representable at original size, and at a higher signal to noise ratio. Interpolation techniques generally blur important edge information in images or videos.

In recent years, deep learning techniques have shown to be very promising for a variety of image and video enhancement tasks, including super-resolution [9][10]. One possible approach to perform video super-resolution is to apply SISR on each frame of a given video. This may result in various artifacts such as flickering effects, if frames of the video are considered independent of each other. Using temporal information is a natural step to preserve temporal consistency. Some recent deep learning methods have applied concatenated frames [15] and some have applied Recurrent Neural Networks (RNNs) to preserve temporal consistency [18].

In this manuscript, we propose an end-to-end trainable video super-resolution method which is an extension of [17][18] that has been recently applied to single image super-resolution. Our method uses temporal information as well as single frame structural information to construct high resolution frames of a video. The frames are temporally consistent with each other and are of higher quality compared to its corresponding frame-wise single image super-resolution. To achieve this goal, we incorporated temporal information using optical flow to track pixels in an adversarial network. The adversarial networks are chosen for this task as they have historically shown better performance in generating sharp and realistic output. To further increase the temporal consistency, we provided an extra condition on the output of the discriminator. We provided the previous frame along with the earlier input to the discriminator to make the output of the discriminator dependent on the previous frame to impose temporal consistency. The major contributions of this paper are as follows.

1. We proposed an adversarial based two-stage network that incorporates temporal information along with the structural information of current frame to generate outputs that are realistic in nature when considered independently and are also temporally consistent when taken as a frame of the given video.
2. Along with providing temporal information to the generator to create sharp, realistic and temporally consistent output, we added an extra condition on the discriminator that ensures temporal consistency.
3. We trained our model on the REDS (REalistic and Dynamic Scenes) dataset and compared our results with the bicubic interpolation and the frame-by-frame single image super-resolution method both quantitatively and qualitatively.

2 Related Work

Super-resolution is an ill-posed inverse problem. Classical interpolation methods such as bicubic interpolation have been traditionally used for resizing a given single image. Single image Super-Resolution (SISR) is also a highly studied problem in the context of deep learning schemes. Deep learning-based SISR was first introduced in SRCNN [2] that requires a predefined upsampling operator. Ledig et al introduced SRGAN [19] that uses a GAN-based framework for generating realistic images [20]. There were other improvements made in SISR using deep learning such as introduction of upsampling layers [21], back-projection [22] and progressive upsampling [4]. Nazeri et al [17] have used an edge-informed two stage network to address the SISR problem as a specific case of an image inpainting problem.

Similar to SISR, video super-resolution is also an extensively studied problem. Earlier traditional methods include [23][24] assumed to be the best transformation among frames. Further, Propper et al. [25] generalized the non-local means framework for video SR in order to handle complex motion patterns in videos.

Since the introduction of SRCNN [2], deep learning methods have also evolved as a tool to address VSR. Most of these techniques include a two step framework such as the one used in [26] by Kappelar et al. Optical flow is first estimated and then compensated frames are fed into a convolutional neural network that constructs a high-resolution frame. Several other methods such as [27] have also used optical flow to estimate relative motion between images. They have then performed warping temporal alignment. The DUF method in [15] has used implicit motion compensation for video super-resolution. EDVR in [28] has used deformable convolutional networks for video super-resolution. Haris et al have used a recurrent architecture for video super-resolution task [22].

3 Proposed Method

We propose a video super-resolution framework that consists of two stages: (i) Edge enhancement, i.e., generation of a high resolution edge map of the given low resolution frame, (ii) Image completion, i.e., generation of a high resolution frame of the given low resolution frame. Similar to [17], both stages have their own adversarial model consisting of a generator-discriminator pair. Let \(g_1\) and \(D_1\) be the generator and discriminator of edge enhancement stage respectively and \(g_2\) and \(D_2\) be the generator and discriminator of the image completion stage. Our method uses the temporal information from the previous frames along with spatial information from the current frame in order to reconstruct high resolution frames that are temporally consistent. The temporal information is provided in the generator and discriminator pair for both edge generation and image completion stages. This modification will be explained in the following sections.

3.1 Incorporating Temporal Information

One way to incorporate temporal information in VSR is to provide the previous frame along with current frame as an input to the model. Given a low resolution frame \(I^L\), we denote its previous frame as \(I^R\). Our goal is to create a reconstructed/predicted high resolution image \(I^H_{pred}\) of the previous low resolution frame to be passed to the network. For the very first frame that has no previous frame, we use its bicubic interpolated high resolution image as its previous frame.

To further enhance the model, in addition to providing the reconstructed high resolution previous frame as a prior, we also calculate and pass the optical flow vector corresponding to the video.
sequence. The calculated optical flow vectors $F_{HR}$ will be computed in between the interpolated versions of low resolution previous and current frame, using bicubic interpolation. We then warp the predicted previous frame $\hat{I}_{pred}$ using the optical flow vector $F_{HR}$ to obtain a warped reconstructed previous frame $\hat{I}_{warp}$. For the very first frame of the video, we assume the optical flow vectors are zero. The process of calculating $\hat{I}_{warp}$ can be summarized as

$$F_{HR} = \text{OpticalFlow} \left( \text{UpSample}(I_{LR}), \text{UpSample}(I_{HR}) \right),$$

$$\hat{I}_{warp} = \text{Warp}(\hat{I}_{pred}, F_{HR}),$$

where $I_{LR}$ and $I_{HR}$ are the current and previous low resolution frames, respectively, for which an optical flow vector is being computed. UpSample is an upsampling module, which is a bicubic interpolation kernel for a zooming factor of 4 in our experiments. Finally, Warp is a forward warping transformation applied to $\hat{I}_{pred}$ given the high resolution flow vector $F_{HR}$.

We provide the warped predicted previous frame $\hat{I}_{warp}$ as a prior by concatenating it with other inputs of the model. All the other inputs relating to the generator and discriminator for both edge generation and image completion stages will be explained in their respective sections. The estimated $\hat{I}_{warp}$ is provided to the generator and the discriminator at both stages. In this fashion, the generator aims to generate an output which is spatially and temporally consistent and the discriminator checks whether the output is spatially and temporally consistent or not.

### 3.2 Edge Enhancement

The high resolution edge map of the current frame is reconstructed at the edge enhancement stage, similar to [17]. Let $I^{LR}$ and $C^{LR}$ denote the current low resolution frame and its corresponding low resolution edge map, respectively. We have also included an additional nearest neighbour module to resize $I^{LR}$ and $C^{LR}$ to the same size as the high resolution image, see Figure 1. The generator of the edge enhancement stage $G$ predicts the high resolution edge map $C_{pred}$ of the current low resolution frame by taking the inputs $I^{LR}, C^{LR}$ and warped reconstructed previous frame $\hat{I}_{warp}$, i.e.,

$$C_{pred} = G(I^{LR}, C^{LR}, \hat{I}_{warp}).$$

The architecture of our proposed edge enhancement stage is shown in Figure 1. The hinge adversarial loss [29] is used in the generator $G$ and discriminator $D$ of the edge enhancement stage, defined similar to [17] as

$$L_{Gi} = -E_{\mathcal{P}_x} \left[ D_i(C_{pred}, I^{LR}, \hat{I}_{warp}) \right],$$

and

$$L_{Di} = E_{(C_{gt}, I^{LR})} \left[ \max(0, 1 - D_i(C_{gt}, I^{LR}, \hat{I}_{warp})) \right] + E_{\mathcal{P}_x} \left[ \max(0, 1 + D_i(C_{pred}, I^{LR}, \hat{I}_{warp})) \right].$$

where $C_{gt}$ represents the ground truth high resolution edge map. We have used a feature matching loss $\mathcal{L}_{fm}$ for our edge enhancement generator. This feature matching loss compares activation maps in the intermediate layers of the discriminator. The constraint here is on the generator to predict and produce results which have edge maps similar to ground-truth high resolution edge maps. The feature matching loss is defined as

$$\mathcal{L}_{fm} = E \left[ \sum_{i} \frac{1}{N_i} \left\| D_i(C_{gt}) - D_i(C_{pred}) \right\|_1 \right].$$

where $N_i$ is the number of elements in the $i^{th}$ activation layer, and $D_i$ is the activation in the $i^{th}$ layer of the discriminator. Spectral normalization (SN) [29] further stabilizes training by scaling down weight matrices by their respective largest singular values. We apply SN to both the generator and discriminator [30, 31]. The final joint loss objective for $G_i$ with regularization parameters $\lambda_G$ and $\lambda_{fm}$ becomes

$$S_{Gi} = \lambda_G L_{Gi} + \lambda_{fm} L_{fm},$$

where we choose $\lambda_G = 1$ and $\lambda_{fm} = 10$ for all experiments.

### 3.3 Image Completion

During the image completion stage, the current low resolution frame $I^{LR}$ is initially converted into an incomplete high resolution frame represented by $I^{HR}$ using a fixed fractionally strided convolution kernel, see Figure 2. This has the effect of adding empty rows and columns in-between pixels. In order to increase the size of $I^{LR}$ by a factor of 4, we used a $4 \times 4$ kernel $K$ with all the values equal to 0 except the top left value which was set to 1. The value of stride was set to 1/4. The incomplete high resolution $I^{HR}$ computed as

$$I^{HR} = I^{LR} \ast K$$

was passed to the image completion network generator $G_2$ along with $C_{pred}$ and $\hat{I}_{warp}$ as inputs to generate the predicted high resolution current frame $\hat{I}_{pred}$, i.e.,

$$I_{pred} = G_2(I^{HR}, C_{pred}, \hat{I}_{warp}).$$

The generator $G_2$ was trained with a combination of hinge loss, $\ell_1$ loss $\mathcal{L}_{c} = \ell_1$, perceptual loss [32], and style loss, similar to [17]. The architecture of our proposed image completion stage is shown in Figure 2. The hinge loss ensures that the images generated by generator are of realistic nature. Similar to the losses given in Equation (4) for image edge generation stage, at the image completion stage we define

$$L_{G_2} = E_{(I_{pred}, C_{pred})} \left[ D_2(I_{pred}, C_{pred}, \hat{I}_{warp}) \right] + E_{(I_{gt}, C_{pred})} \left[ \max(0, 1 - D_2(I_{gt}, C_{pred}, \hat{I}_{warp})) \right],$$

$$L_{D_2} = E_{(I_{gt}, C_{pred})} \left[ \max(0, 1 - D_2(I_{gt}, C_{pred}, \hat{I}_{warp})) \right] + E_{(I_{pred}, C_{pred})} \left[ \max(0, 1 + D_2(I_{pred}, C_{pred}, \hat{I}_{warp})) \right].$$
In our experiments, we used parameters $\lambda_0$, $\beta_1$, and $\beta_2$ that seemed to be effective for the video super-resolution task. The schematic diagram of the proposed method of generating high resolution images is shown in Figure 3.

While the perceptual loss $L_{\text{perc}}$ encourages perceptual similarity between generated and ground truth images, the style loss $L_{\text{style}}$ tends to maintain the texture similarity by minimizing the Euclidean distance between the Gram matrices of the intermediate feature maps. The style loss is defined as

$$L_{\text{style}} = \mathbb{E} \left[ \sum_j \| G_j^b(I_g) - G_j^b(I_{\text{pred}}) \|_1 \right],$$  

in which $G_j^b$ represents the Gram matrix constructed from activation maps $\phi_j$. The main purpose to adding the style loss was to mitigate the “checkerboard” artifact caused by transpose convolutions as shown by Sajjadi et al. For both style and perceptual loss we extract feature maps from $\text{relu}11$, $\text{relu}21$, $\text{relu}31$, $\text{relu}41$ and $\text{relu}51$ of VGG-19. The proposed total combined objective for the image completion stage is

$$3G_5 = \lambda_1 L_{t_i} + \lambda_2 L_{G_5} + \lambda_p L_{\text{perc}} + \lambda_s L_{\text{style}}.$$  

In our experiments, we used parameters $\lambda_1 = 1$, $\lambda_2 = \lambda_p = 0.1$, and $\lambda_s = 250$ that seemed to be effective for the video super-resolution task. The schematic diagram of the proposed method of generating HR video from LR video is shown in Figure 3.

### 4 Experimental Results

#### 4.1 Dataset

We used REDS (REalistic and Dynamic Scenes) data which is a high quality (720p) video dataset, for the training and testing. REDS training contains 24,000 frames taken from 240 videos, while validation set contains 30,000 frames taken from 30 videos. Each video has 100 frames of $1280 \times 720 \times 3$ size. We have selected 4 videos (000, 011, 015, 020) from validation and used them for testing purpose. This dataset is also known as REDS_S4 dataset. The rest of the videos from training and validation set are used for training.

#### 4.2 Training Details

For the edge generation stage relating to generator $G_1$ we used Canny edge detection. Its hyper-parameters are minimum and maximum Hysteresis thresholding value and Sobel kernel size. We chose values 100, 200 and 3 for the former mentioned parameters respectively. We used Adam optimizer for the generator and discriminator at both stages with a learning rate of 0.00001. The value of $\beta_1$ and $\beta_2$ were 0.9 and 0.99 respectively. We trained the model for both stages separately. For the edge enhancement stage we trained our model for 21 epochs. For the image completion stage, our model was trained for 15 epochs.
4.2.1 Two-step Training for Improved Stability

Our model required warped reconstructed previous frame as a prior. At the very first step that no reconstructed frames are available, we can use bicubic interpolation to obtain an estimate of high resolution frames. We notice that during the initial training, the reconstructed images usually appear as random noise. If we use these poor quality images as a reconstructed frame and pass them as a prior after warping them using optical flow, the training will be unstable. To handle this, we first used actual ground truth high resolution previous frames for the warping and provided these warped actual high resolution previous frames as a prior. After some epochs when the model has been trained sufficiently to generate good quality high resolution frames, we started training the model using warped reconstructed previous frame as a prior.

4.2.2 Qualitative Evaluation

The qualitative comparison of our method along with Histogram matching (HM) compared to the bicubic interpolation and frame-by-frame edge-informed single image super-resolution (SISR) method [17] and SISR [17] + HM on REDS_S4 dataset is presented in Figure 4. We can observe that the displayed video frames of our proposed method in Figure 4 are superior and closer to ground truth in terms of quality, sharpness, and are more realistic in nature. All of the video results are accessible at: https://bit.ly/2HWaIHT

4.2.3 Quantitative Evaluation

The quantitative comparison of our proposed method along with Histogram matching (HM) compared to bicubic interpolation and frame-by-frame edge-informed single image super-resolution (SISR) method [17] & SISR [17] + HM on REDS_S4 dataset is presented in Tables 1 and 2 with respect to SSIM and PSNR.

Table 1: Quantitative evaluation with respect to SSIM of various methods with our proposed model on REDS_S4 dataset

<table>
<thead>
<tr>
<th>Method, Video #</th>
<th>000</th>
<th>011</th>
<th>015</th>
<th>020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>0.6489</td>
<td>0.7261</td>
<td>0.8034</td>
<td>0.7386</td>
</tr>
<tr>
<td>SISR [17]</td>
<td>0.6607</td>
<td>0.7278</td>
<td>0.8113</td>
<td>0.7186</td>
</tr>
<tr>
<td>SISR [17] + HM</td>
<td>0.6676</td>
<td>0.7469</td>
<td>0.8180</td>
<td>0.7258</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.7426</td>
<td>0.7469</td>
<td>0.8562</td>
<td>0.7908</td>
</tr>
<tr>
<td>Proposed + HM</td>
<td>0.7460</td>
<td>0.8104</td>
<td>0.8553</td>
<td>0.7922</td>
</tr>
</tbody>
</table>

Table 2: Quantitative evaluation with respect to PSNR of various methods with our proposed model on REDS_S4 dataset

<table>
<thead>
<tr>
<th>Method, Video #</th>
<th>000</th>
<th>011</th>
<th>015</th>
<th>020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>24.55</td>
<td>26.06</td>
<td>28.52</td>
<td>25.41</td>
</tr>
<tr>
<td>SISR [17]</td>
<td>22.76</td>
<td>22.87</td>
<td>27.28</td>
<td>23.03</td>
</tr>
<tr>
<td>Ours</td>
<td>24.35</td>
<td>25.13</td>
<td>28.87</td>
<td>24.74</td>
</tr>
<tr>
<td>Ours + HM</td>
<td>25.73</td>
<td>26.93</td>
<td>29.64</td>
<td>25.66</td>
</tr>
</tbody>
</table>

5 Conclusions

We proposed a novel Edge-VSR method that includes a two-stage trainable network able to generate high quality results with temporal consistency. In extensive experiments we have shown that our proposed method outperformed the baseline i.e. frame-by-frame edge informed SISR [17] and bicubic interpolation.

Acknowledgements

This research was supported by an NSERC-DDG & was completed during remote internship of A.S. at Ontario Tech.
References


