Temporally Consistent Edge-Informed Video Super-Resolution (Edge-VSR)

Ayush Singh
Indian Institute of Technology
Dhanbad, Jharkhand, India
ayush.s.18je0204@cse.iitism.ac.in

Mehran Ebrahimi
Imaging Lab, Faculty of Science, Ontario Tech University
Oshawa, Ontario, Canada
mehran.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/2HWaIHT

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 VSR methods have applied concatenated frames [15] and some have applied Recurrent Neural Networks (RNNs) to preserve temporal consistency [16].

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] assume an affine transformation exist between adjacent frames. Further, Protter 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^{LR}, we denote its previous frame as $\hat{I}^{\mathrm{LR}}$. Our goal is to create a reconstructed/predicted high resolution image $\hat{I}_{pred}$ 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

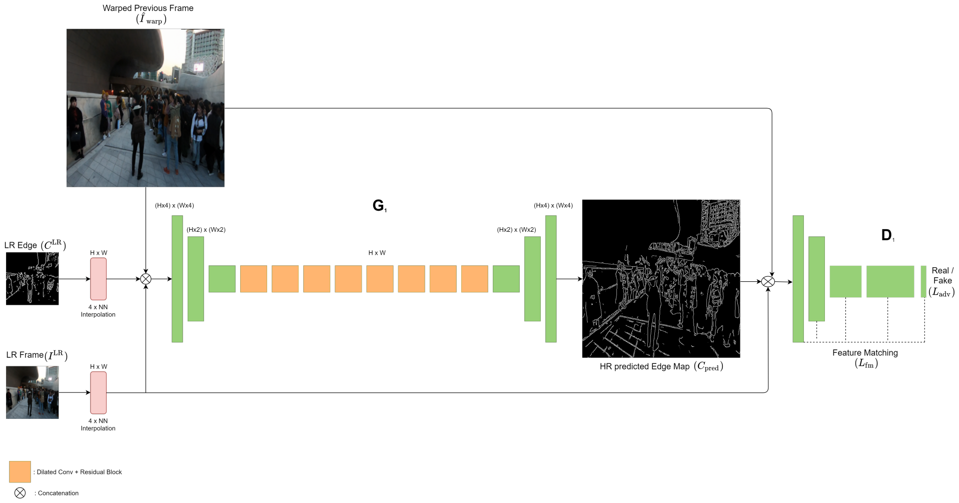
Warped Previous Frame
$(\hat{I}_{\text{warp}})$
LR Edge (C^{LR})
H x W
4 x NN Interpolation
LR Frame (I^{LR})
H x W
4 x NN Interpolation
(Hx4) x (Wx4)
(Hx2) x (Wx2)
G_1
H x W
(Hx2) x (Wx2)
(Hx4) x (Wx4)
HR predicted Edge Map (C_{pred})
D_1
Real / Fake (L_{adv})
Feature Matching (L_{fm})
: Dilated Conv + Residual Block
: Concatenation

Fig. 1: Schematic architecture of the proposed edge enhancement stage

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

$$F^{\text{HR}} = \text{OpticalFlow}\Big(\text{UpSample}(I^{\text{LR}}), \text{UpSample}(\hat{I}^{\text{LR}})\Big), \tag{1}$$

$$\hat{I}_{warp} = \text{Warp}(\hat{I}_{pred}, F^{\text{HR}}), \tag{2}$$

where I^{LR} and $\hat{I}^{\text{LR}}$ are the current and previous low resolution frames, respectively, for which an optical flow vector is being computed. UpSample is an upscaling 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_1 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.,

$$C_{\text{pred}} = G_1(I^{\text{LR}}, C^{\text{LR}}, \hat{I}_{warp}). \tag{3}$$

The architecture of our proposed edge enhancement stage is shown in Figure 1. The hinge adversarial loss [29] is used in the generator G_1 and discriminator $D1$ of the edge enhancement stage, defined similar to [17] as

$$\mathcal{L}_{\text{G}_1} = -\mathbb{E}_{I^{\text{LR}}}\Big[D_1(C_{\text{pred}}, I^{\text{LR}}, \hat{I}_{warp})\Big], \tag{4}$$

and

$$\mathcal{L}_{\text{D}_1} = \mathbb{E}_{(C_{\text{gt}}, I^{\text{LR}})}\Big[\max(0, 1 - D_1(C_{\text{gt}}, I^{\text{LR}}, \hat{I}_{warp}))\Big] + \mathbb{E}_{I^{\text{LR}}}\Big[\max(0, 1 + D_1(C_{\text{pred}}, I^{\text{LR}}, \hat{I}_{warp}))\Big],$$

where C_{gt} represents the ground truth high resolution edge map. We have used a feature matching loss $\mathcal{L}_{\text{fm}}$ 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

$$\mathcal{L}_{\text{fm}} = \mathbb{E}\Big[\sum_i \frac{1}{N_i} ||D_1^{(\text{i})}(C_{gt}) - D_1^{(\text{i})}(C_{pred})||_1\Big], \tag{5}$$

where N_i is the number of elements in the i^{th} activation layer, and $D_1^{(\text{i})}$ 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_1 with regularization parameters λ_{G_1} and λ_{FM} becomes

$$\Im_{\text{G}_1} = \lambda_{\text{G}_1}\mathcal{L}_{\text{G}_1} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}, \tag{6}$$

where we choose λ_{G_1} = 1 and λ_{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 $\dot{I}^{\text{HR}}$ 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 4×4 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 $\dot{I}^{\text{HR}}$ computed as

$$\dot{I}^{\text{HR}} = I^{\text{LR}} * K \tag{7}$$

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 I_{pred}, i.e.,

$$I_{\text{pred}} = G_2(\dot{I}^{\text{HR}}, C_{\text{pred}}, \hat{I}_{warp}). \tag{8}$$

The generator G_2 was trained with a combination of hinge loss, ℓ_1 loss $\mathcal{L}_{\text{fm}}$, 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

$$\mathcal{L}_{\text{G}_2} = \mathbb{E}_{C_{\text{pred}}}\Big[D_2(I_{\text{pred}}, C_{\text{pred}}, \hat{I}_{warp})\Big], \tag{9}$$

$$\mathcal{L}_{\text{D}_2} = \mathbb{E}_{(I_{\text{gt}}, C_{\text{pred}})}\Big[\max(0, 1 - D_2(I_{\text{gt}}, C_{\text{pred}}, \hat{I}_{warp}))\Big] + \mathbb{E}_{C_{\text{pred}}}\Big[\max(0, 1 + D_2(I_{\text{pred}}, C_{\text{pred}}, \hat{I}_{warp}))\Big],$$

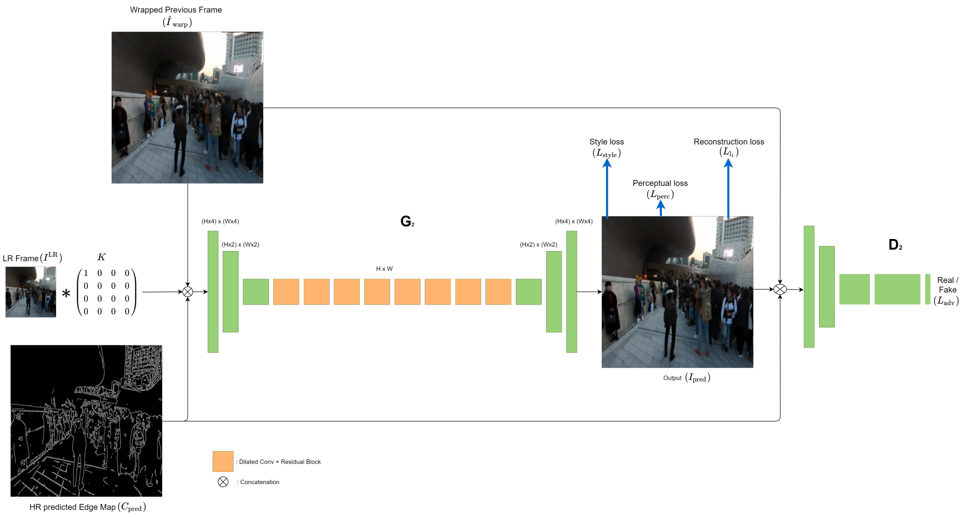
Wrapped Previous Frame
($\hat{I}_{\text{warp}}$)
LR Frame (I^{LR})
K
$* \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$
HR predicted Edge Map (C_{pred})
(Hx4) x (Wx4)
(Hx2) x (Wx2)
G₂
H x W
(Hx2) x (Wx2)
(Hx4) x (Wx4)
Style loss
(L_{style})
Perceptual loss
(L_{perc})
Reconstruction loss
(L_{l_1})
Output (I_{pred})
D₂
Real / Fake
(L_{adv})
: Dilated Conv + Residual Block
: Concatenation

Fig. 2: The architecture of the proposed image completion stage

in which I_{gt} is the ground-truth high resolution frame.

The perceptual loss $\mathcal{L}_{\text{perc}}$ that we used in our objective minimizes the L1 distance between feature maps generated from intermediate layers of VGG-19 trained on the ImageNet dataset [33]. This loss provides an additional constraint on the generator to produce high resolution images that are perceptually similar to the ground truth. The perceptual loss is defined as

$$\mathcal{L}_{\text{perc}} = \mathbb{E}\Big[\sum_i \frac{1}{N_i} || \ \phi_i(I_{gt}) - \phi_i(I_{pred}) \ ||_1\Big], \tag{10}$$

where N_i is the number of elements in the i^{th} activation of VGG-19. Expressions $\phi_i(I_{gt})$ and $\phi_i(I_{pred})$ represent the feature maps of I_{gt} and I_{pred} respectively, corresponding to the i^{th} activation layer of VGG-19. The feature matching loss also encourages the perceptual similarity between the generated and ground truth images but perceptual loss is shown to be much effective for image generation task and adding feature matching loss along with perceptual loss might be redundant [32–34].

While the perceptual loss $\mathcal{L}_{\text{perc}}$ encourages perceptual similarity between generated and ground images, the style loss $\mathcal{L}_{\text{style}}$ tends to maintain the texture similarity by minimizing the ℓ_1 distance between the Gram matrices of the intermediate feature maps. The style loss is defined as

$$\mathcal{L}_{\text{style}} = \mathbb{E}\Big[\sum_j || \ G_j^{\phi}(I_{gt}) - G_j^{\phi}(I_{pred}) \ ||_1\Big], \tag{11}$$

in which G_j^{ϕ} represents the Gram matrix constructed from activation maps ϕ_j [17]. The main purpose to adding the style loss was to mitigate the“checkerboard” artifact caused by transpose convolutions [35] as shown by Sajjadi et al. [34]. For both style and perceptual loss we extract feature maps from $relu11$, $relu21$, $relu31$, $relu41$ and $relu51$ of VGG-19. The proposed total combined objective for the image completion stage is

$$\Im_{\text{G}_2} = \lambda_{l_1}\mathcal{L}_{\ell_1} + \lambda_{\text{G}_2}\mathcal{L}_{\text{G}_2} + \lambda_{\text{p}}\mathcal{L}_{\text{perc}} + \lambda_{\text{s}}\mathcal{L}_{\text{style}}. \tag{12}$$

In our experiments, we used parameters λ_{ℓ_1} = 1, λ_{G_2} = λ_{p} = 0.1, and λ_{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.

3.4 Histogram Matching

In our proposed method, we have added a histogram matching (HM) step as additional post-processing step. After the high resolution frame I_{pred} is generated by G_2, we perform histogram matching (HM) on I_{pred}, to match the histogram of the bicubic interpolated copy of the low resolution current frame that has the same size as the generated high resolution frame. In our experiments, we realized that this post processing step increased quality of the results.

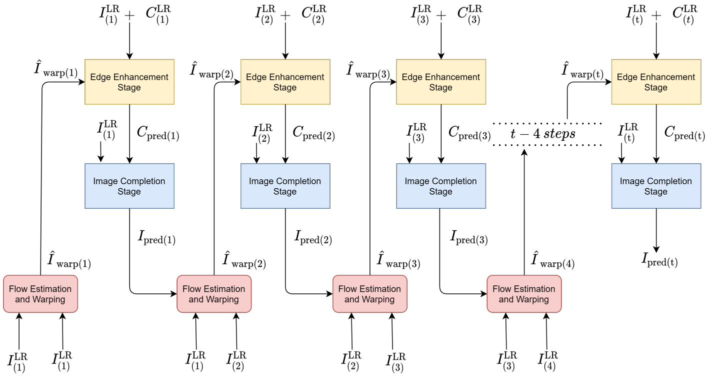

Fig. 3: Schematic diagram of the proposed method. Note that for every frame number t in a video, $I^{\text{LR}}_{(t)}$ represents t^{th} low resolution frame, $C^{\text{LR}}_{(t)}$ is the low resolution edge map, $\hat{I}_{\text{warp}}(t)$ is the warped high resolution previous frame $(t-1)$, finally $I_{\text{pred}}(t)$ and $C_{\text{pred}}(t)$ are the predicted high resolution t^{th} frame and its corresponding edge map, respectively.

4 Experimental Results

4.1 Dataset

We used REDS (REalistic and Dynamic Scenes) data [36] 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 [37] for the generator and discriminator at both stages with a learning rate of 0.00001. The value of β_1 and β_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.

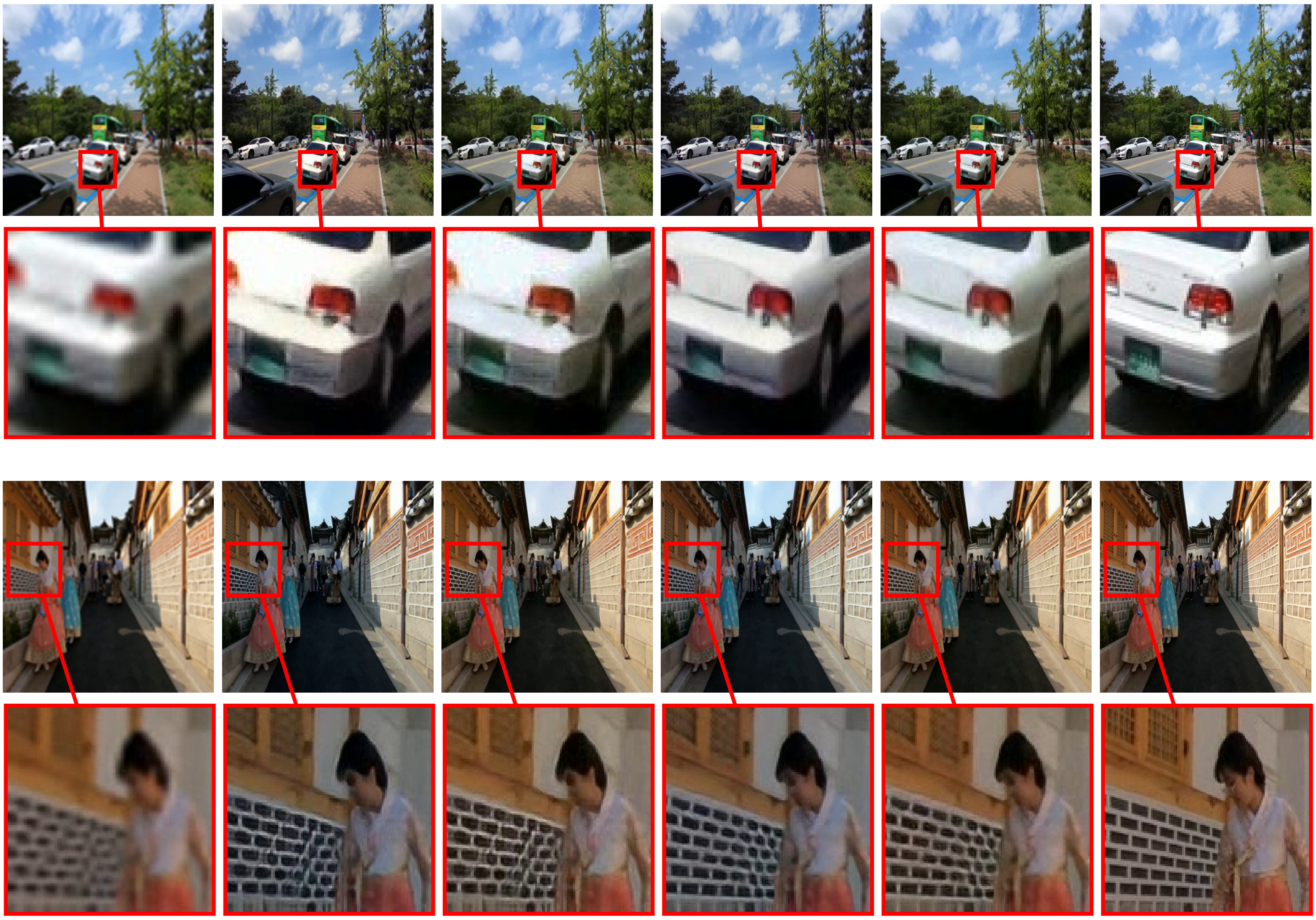

Fig. 4: Qualitative comparison of our model with respect to other methods on REDS_S4 dataset, from left to right Bicubic, SISR [17], SISR [17]+ HM, Proposed, Proposed + HM, and Ground Truth, [HM stands for Histogram Matching]

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 RED_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/2HWalHT

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 RED_S4 dataset is presented in Tables 1 and 2 with respect to SSIM and PSNR.

Method, **Video #**	**000**	**011**	**015**	**020**
Bicubic	0.6489	0.7261	0.8034	0.7386
SISR [17]	0.6607	0.7278	0.8113	0.7186
SISR [17] + HM	0.6676	0.7469	0.8180	0.7258
Proposed	0.7426	0.7469	0.8562	0.7908
Proposed + HM	0.7460	0.8104	0.8553	0.7922

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

Method, **Video #**	**000**	**011**	**015**	**020**
Bicubic	24.55	26.06	28.52	25.41
SISR [17]	22.76	22.87	27.28	23.03
SISR [17] + HM	23.90	25.01	28.36	23.99
Ours	24.35	25.13	28.87	24.74
Ours + HM	25.73	26.93	29.64	25.66

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

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

[1] M. Haris, M. R. Widyanto, and H. Nobuhara, “Inception learning super-resolution,” *Applied optics*, vol. 56, no. 22, pp. 6043–6048, 2017.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[3] M. Ebrahimi and E. R. Vrscay, “Regularization schemes involving self-similarity in imaging inverse problems,” in *Proceedings of Applied Inverse Problems (AIP), DOI:10.1088/1742-6596/124/1/012021, 12 pages*, University of British Columbia, Vancouver, Canada, June 2007.

[4] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[5] M. Ebrahimi and E. R. Vrscay, “Solving the inverse problem of image zooming using “self-examples”,” in *Lecture Notes in Computer Science, Proceedings of The International Conference on Image Analysis and Recognition ICIAR*, vol. 4633. Montreal, Canada: Springer, August 2007, pp. 117–130.

[6] M. Ebrahimi and S. Bohun, “Single image super-resolution via non-local normalized graph laplacian regularization: A self-similarity tribute,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 93, p. 105508, 2021.

[7] M. Ebrahimi and E. R. Vrscay, “Nonlocal-means single-frame image zooming,” in *Proceedings in Applied Mathematics and Mechanics (PAMM), 6th International Congress on Industrial and Applied Mathematics, ICIAM*, vol. 7, no. ’1, Zurich, Switzerland, July 2007, pp. 2 020 067–2 020 068.

[8] D. C. Garcia, C. Dorea, and R. L. de Queiroz, “Super resolution for multiview images using depth information,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1249–1256, 2012.

[9] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, “Deep video super-resolution using hr optical flow estimation,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4323–4336, 2020.

[10] M. Haris, G. Shakhnarovich, and N. Ukita, “Recurrent back-projection network for video super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3897–3906.

[11] M. S. Sajjadi, R. Vemulapalli, and M. Brown, “Frame-recurrent video super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6626–6634.

[12] M. Ebrahimi and A. L. Martel, “A PDE approach to coupled super-resolution with non-parametric motion,” in *Lecture Notes in Computer Science, Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition (EMM-CVPR)*, D. Cremers, Y. Boykov, A. Blake, and F. R. Schmidt, Eds., vol. 5681. Bonn, Germany: Springer-Verlag, August 2009, pp. 112–125.

[13] M. Ebrahimi, E. R. Vrscay, and A. L. Martel, “Coupled multi-frame super-resolution with diffusive motion model and total variation regularization,” in *Proceedings of The International Workshop on Local and Non-Local Approximation in Image Processing (LNLA)*, J. Astola, K. Egiazarian, and V. Katkovnik, Eds. Tuusula, Finland: Tampere International Center for Signal Processing, August 2009, pp. 62–69.

[14] M. Ebrahimi and E. R. Vrscay, “Multi-frame super-resolution with no explicit motion estimation,” in *Proceedings of The International Conference on Image Processing, Computer Vision, and Pattern Recognition, IPCV*, Las Vegas, Nevada, USA, July 2008, pp. 455–459.

[15] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, “Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3224–3232.

[16] Y. Huang, W. Wang, and L. Wang, “Bidirectional recurrent convolutional networks for multi-frame super-resolution,” in *Advances in Neural Information Processing Systems*, 2015, pp. 235–243.

[17] K. Nazeri, H. Thasarathan, and M. Ebrahimi, “Edge-informed single image super-resolution,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[18] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, “Edgeconnect: Generative image inpainting with adversarial edge learning,” in *Proceedings of ICCV Workshops (Advances in Image Manipulation), https://arxiv.org/abs/1901.00212*, Seoul, Korea, November 2019.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[21] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[22] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-projection networks for super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.

[23] R. R. Schultz and R. L. Stevenson, “Extraction of high-resolution frames from video sequences,” *IEEE transactions on image processing*, vol. 5, no. 6, pp. 996–1011, 1996.

[24] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, “Joint map registration and high-resolution image estimation using a sequence of undersampled images,” *IEEE transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[25] M. Protter, M. Elad, H. Takeda, and P. Milanfar, “Generalizing the nonlocal-means to super-resolution reconstruction,” *IEEE Transactions on image processing*, vol. 18, no. 1, pp. 36–51, 2008.

[26] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, “Video super-resolution with convolutional neural networks,” *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[27] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, “Real-time video super-resolution with spatio-temporal networks and motion compensation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4778–4787.

[28] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, “Edvr: Video restoration with enhanced deformable convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” *arXiv preprint arXiv:1802.05957*, 2018.

[30] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[31] A. Odena, J. Buckman, C. Olsson, T. B. Brown, C. Olah, C. Raffel, and I. Goodfellow, “Is generator conditioning causally related to gan performance?” *arXiv preprint arXiv:1802.08768*, 2018.

[32] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[34] M. S. Sajjadi, B. Scholkopf, and M. H. EnhanceNet, “Single image super-resolution through automated texture synthesis,” *Max-Planck-Institute for Intelligent Systems Spemanstr*, vol. 23, 2016.

[35] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts. distill (2016),” 2016.

[36] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. Mu Lee, “Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[37] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.